Predictive Crossvalidation and Baseline Correction in Mixed Models for Longitudinal Data

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Julia Braun

aus

Deutschland

Promotionskomitee

Prof. Dr. Leonhard Held (Vorsitz)
Prof. Dr. Reinhard Furrer
Prof. Dr. Peter Bühlmann (ETH Zürich)

Zürich, 2013

Preface

Many people have helped and supported me during the work on my PhD thesis. Now it is time to express my sincere gratitude.

First of all, I would like to thank my supervisor Leonhard Held for many very helpful discussions, his continuing interest in my work, his loyalty and great patience. He was always open for questions and encouraged me to attend several interesting conferences. I also thank the members of my dissertation committee, Reinhard Furrer and Peter Bühlmann, as well as my external referee Thomas Kneib for reviewing my thesis.

I am very grateful to all my colleagues at the Division of Biostatistics for their support. I especially want to mention Michaela Paul, Andrea Riebler, Daniel Sabanés Bové and Birgit Schrödle, who always encouraged me and tried to answer my questions, no matter how stupid or difficult they were or how much time was needed. Special thanks go to Sarah Haile who shared an office with me and was always there to comfort me when things didn't work well, who proofread the introduction and parts of the papers, and who showed interest in every detail I wanted to talk about.

I also want to thank Matthias Bopp and David Fäh, my collaborators from the Unit of Demography and Health Statistics, for giving me the opportunity to work in their department while still occupied with my dissertation. They were always very supportive and understanding. I owe thanks to Bruno Ledergerber from the Division of Infectious Diseases and Hospital Epidemiology at the University Hospital Zurich for many explanations concerning HIV.

Moreover, I am thankful to all my friends in and outside the world of statistics for being there for me when I need somebody to talk to. I am especially obliged to Jürgen Deinhard who helped me with two proofs and always enjoyed talking about the daily life of a biostatistician.

Additionally, I would like to thank my parents, Dr. Isabel Grübel and Rainer Possmann, and my grandparents, Dr. Hans and Ingrid Grübel, for their continuous love and support. Although not familiar with statistical issues, they were always interested in my work and helped whenever possible. Many thanks go especially to Rainer Possmann for proofreading parts of the manuscript.

Finally, I want to thank my husband Sebastian and my son Severin for going all this long way with me. Especially during the last year, I was not able to spend as much time with them as I would have liked. They were always loving and understanding and lifted my spirits by just being there for me. Thank you so much!

Zurich, December 2012

Julia Braun

Zusammenfassung

Gemischte Modelle bilden eine sehr flexible Klasse von Modellen zur Analyse longitudinaler Daten. Sowohl feste Effekte, die sich auf die Gesamtpopulation beziehen, als auch individuelle zufällige Effekte können so geschätzt werden. Zusätzlich kann serielle Korrelation verwendet werden, die Abhängigkeiten zwischen Messungen desselben Individuums über die Zeit berücksichtigt. Aufgrund der speziellen Struktur longitudinaler Daten ist es aber leider in vielen Fällen nicht möglich, bestimmte Methoden anzuwenden, die bei normalen linearen Modellen relativ einfach sind. Stattdessen müssen diese Methoden für die Verwendung in gemischten Modellen angepasst werden. Zwei Beispiele, bei denen substantielle Veränderungen bestimmter Methoden nötig sind, wenn sie bei gemischten Modellen verwendet werden sollen, werden in dieser Arbeit diskutiert.

Zunächst beschäftigen wir uns mit dem Problem der Modellwahl. In normalen linearen oder generalisierten linearen Modellen müssen nur die Einflussgrössen gewählt werden. In gemischten Modellen ist jedoch auch eine Entscheidung bezüglich der Berücksichtigung von zuälligen Effekten und serieller Korrelation nötig. Übliche Modellwahlkriterien wie Akaikes Informationskriterium (AIC) und das Baysianische Informationskriterium (BIC) müssen zu diesem Zweck verändert werden. Wir schlagen einen alternativen Ansatz zur Wahl linearer gemischter Modelle aus prädiktiver Perspektive vor, wo der Durchschnitt von korrekten Bewertungsregeln, wie dem logarithmischen Score oder dem "continuous ranked probability score", zum Vergleich der Vorhersageeigenschaften verschiedener Modelle dient. Die Verwendung eines Leave-One-Out-Kreuzvalidierungsansatzes, bei dem das jeweilige Modell nur einmal berechnet werden muss, ermöglicht vergleichsweise schnelle Berechnungen. Der Zusammenhang zwischen dem durchschnittlichen kreuzvalidierten logarithmischen Score und dem bedingten AIC wird erläutert, und die Methodik wird anhand eines Datensatzes der Swiss HIV Cohort Study (SHCS) demonstriert mit dem Ziel, ein geeignetes Modell zur Vorhersage der CD4+-Lymphozytenzahlen bei HIV-Patienten zu finden.

In einem zweiten Schritt wird die prädiktive Kreuzvalidierung für die Verwendung bei generalisierten gemischten Modellen erweitert. Dieser Ansatz ist sehr ähnlich wie bei linearen gemischten Modellen und basiert auch auf Kreuzvalidierung mit nur einer Modellanpassung. Allerdings kann hier die prädiktive Verteilung nicht mehr analytisch hergeleitet werden. Daher schlagen wir vor, einen Bayesianischen iterativen gewichteten Kleinste-Quadrate-Algorithmus zur Schätzung der individuellen zufälligen Effekte zu verwenden. Wir demonstrieren die Anwendung dieser Methodik für binär-logistische und log-lineare Poisson-Regression und vergleichen die Ergebnisse mit denen alternativer Methoden.

Zuletzt untersuchen wir, wie man Veränderungen über die Zeit in verschiedenen Gruppen vergleichen kann. Um gültige Vergleiche durchzuführen, muss sichergestellt sein, dass die Veränderung in allen Gruppen bezüglich ähnlicher Startwerte betrachtet wird. Besonders in Beobachtungsstudien sind Messungen zusätzlich auch noch mit Messfehlern behaftet, so dass der wahre Startwert gar nicht beobachtet werden kann. In einem vor kurzem veröffentlichten Artikel wird vorgeschlagen, dieses Problem dadurch zu lösen, dass man ein lineares gemischtes Modell an alle Daten inklusive der Startwerte anpasst und danach die erwartete Veränderung bedingt auf den zugrunde liegenden wahren Startwert berechnet. Da diese Methodik nur eine sehr eingeschränkte Auswahl von Modellen erlaubt, erweitern wir sie, so dass auch zeitabhängige Einflussgrössen und beliebige Interaktionen verwendet werden können. Zusätzlich leiten wir die bedingte erwartete Veränderung in bivariaten Modellen her, so dass auch der Messfehler in anderen zeitvariierenden Einflussgrössen berücksichtigt werden kann. Wir wenden die vorgeschlagene Technik an, um zu zeigen, dass eine gleichzeitige Infektion mit HIV-1 und Hepatitis C eine unterschiedliche Entwicklung der CD4+ Lymphozyten verursacht.

Abstract

Mixed models represent a very flexible and commonly used model class for the analysis of longitudinal data. They allow for the estimation of both population-specific fixed effects and individual random effects. Additionally, serial correlation can be added to cover dependencies of the measurements of the same individual. Unfortunately, the special structure of longitudinal data makes the use of some fairly simple techniques used in normal linear or generalized linear models impossible, and much more refined methods have to be applied. Two examples of such methods that require substantial modifications when intended for mixed models are given in this thesis.

The first issue concerns model choice in mixed models. In a normal linear or generalized linear model, only the covariates have to be chosen. In mixed models, however, a decision on the inclusion and the type of random effects and serial correlation has to be made. Widely used criteria for model choice such as Akaike's information criterion (AIC) or the Bayesian information criterion (BIC) have to be adapted for this task. We present an alternative approach to selection of linear mixed models from a predictive point of view, where mean proper scoring rules like the logarithmic score or the continuous ranked probability score are calculated to assess and compare a model's predictive abilities. An approximate leave-one-out crossvalidation approach where the model has to be fitted just once enables fast computations in comparison to a full leave-one-out crossvalidation. Relations of the mean crossvalidated logarithmic score and the recently proposed conditional AIC are discussed. The methodology is applied to a data set from the Swiss HIV Cohort Study (SHCS) to select a suitable model for predicting the course of CD4+ lymphocyte counts.

Subsequently, the predictive crossvalidation method is extended to the case of generalized linear mixed models. As in the linear mixed model case, the idea of approximate crossvalidation with one single model fit is applied. However, the calculation of the leave-one-out predictive distribution can no longer be done analytically. Therefore, we propose to use a Bayesian iteratively weighted least squares (IWLS) algorithm for the calculation of the individual random effects. Two applications of the methodology for binary logistic and log-linear Poisson regression are presented, and comparisons to alternative methods are shown.

The second issue concerns the comparison of temporal changes in different groups. For valid comparisons, it has to be made sure that changes are compared with respect to similar baseline values in all groups. Especially in observational studies, measurements are subject to measurement error, so that the true baseline value cannot be known. In a recent paper, it is suggested to tackle this problem by fitting a linear mixed model to all data including the baseline measurement, and then calculating the expected change from baseline conditional on the underlying true baseline value. As the original methodology can only be used for a very narrow set of models, we extend it so that time-dependent covariates and arbitrary interactions can be included. Additionally, we derive the expected change from underlying baseline in bivariate models, so that the baseline measurement error of other time-varying covariates is taken into account. In the application, we demonstrate that a joint infection with HIV-1 and hepatitis C leads to different change in CD4+ counts.

Thesis outline

Introduction

Paper I: **Predictive crossvalidation for the choice of linear mixed-effects models with application to data from the Swiss HIV Cohort Study**
Julia Braun, Leonhard Held & Bruno Ledergerber
Paper published in *Biometrics*, 2012, **68** (1), 53-61.

Paper II: **Choice of generalized linear mixed models using predictive crossvalidation**
Julia Braun, Daniel Sabanés Bové & Leonhard Held
Paper under review for *Computational Statistics and Data Analysis*.

Paper III: **Accounting for baseline differences and measurement error in the analysis of change over time**
Julia Braun, Leonhard Held, Bruno Ledergerber & the Swiss HIV Cohort Study
Second revised version under review for *Statistics in Medicine*.

Introduction

Some standard problems that frequently occur in statistical modelling are much more complicated if more than one measurement per individual is available. The particular structure of longitudinal data requires special attention and substantial changes in the methods used for independent observations. Mixed models are a flexible and frequently applied tool for modelling longitudinal data, as they are able to estimate both population effects and characteristics of the individual. But at the same time, several methods known from normal mixed models have to be adapted.

This dissertation tries to tackle two distinct problems that occur when longitudinal data are analyzed: Model choice and the analysis of change over time. In contrast to models for independent observations, model choice incorporates not only the selection of potentially influential covariates, but also of random effects and/or a serial correlation structure. Therefore, specially adapted criteria for model choice are necessary. The second problem discussed in this thesis only occurs when more than one measurement per person is available: When change over time shall be compared for several groups, one has to make sure that this comparison is conducted for groups with the same baseline measurement. This task is made even more difficult in observational studies, because (baseline) measurements are often subject to measurement error, so that the true value can not be observed.

In this introduction, some basic concepts of the models and techniques used in this thesis are explained. In Section 1 the basic structure of linear mixed models for continuous longitudinal data is presented, along with information on estimating those models. Extensions to generalized linear mixed models (GLMMs) and their estimation are discussed in Section 2. Strategies for model choice in general and with special focus on mixed models are introduced in Section 3, followed by an overview of the most common proper scoring rules for performing predictive model validation and criticism in Section 4.

1 Linear mixed models for longitudinal data

Longitudinal data sets consist of repeated measurements of the same individuals, and as such enable to track changes in the outcome of interest. Anticipating the measurements of the same subject to be correlated is the logical and natural consequence of this study design. Therefore, assumptions on the correlation structure of the respective data are mandatory whenever longitudinal data shall be analysed.

There are several possibilities to deal with this need for knowledge on the correlation structure, among them a relatively easy approach where it is not necessary to specify the correlation structure explicitely, and which allows a robust estimation of the covariance matrix (for more information see Diggle *et al.*, 2002, ch. 4). However, this approach ist only recommended for balanced data (i.e. data with the same timepoints and number of measurements for each individual) with short and complete sequences. In the case of unbalanced data, a parametric specification of the covariance structure is strongly recommended.

Some potential sources of random variation can be incorporated in the correlation structure of parametric models for continuous longitudinal data. Among the most likely and intuitive of these sources are serial correlation, random effects and the measurement error. Serial correlation accounts for dependencies between measurements of the same individual due to an underlying time-varying stochastic process. The correlation of two distinct measurements of the same subject depends on the amount of time lying between the two respective dates

where the measurements were taken. In contrast to serial correlation, random effects are not time-dependent and represent the idea of parameters that reflect subject-specific capacities. Therefore they take into account the stochastic variation between different individuals and are a means of controlling this heterogeneity.

Finally, in most situations the existence of some kind of measurement error seems to be a plausible reason for observed variability and should be regarded as a component of the correlation structure. All the three potential sources of random variation mentioned here can be incorporated in parametric models. In the following subsection, we define a flexible parametric model that is able to incorporate all three sources of random variation. Depending on the situation, it can be sufficient to include only one or two of these sources in a particular model. For example, according to Diggle *et al.* (2002, p. 91), it is relatively common that the combination of random effects and measurement error is responsible for most of the observed variation and thus dominates the effects of serial correlation.

1.1 Model formulation

Unlike random effects which are associated with intrinsic properties of the respective individual, fixed effects represent common features of the whole population. Models that combine such fixed and random effects as well as the measurement error are called mixed-effects models or simply mixed models. Apart from the classical linear model, mixed models are one of the most common model classes for continuous longitudinal data. In most situations only one level of grouping is necessary, i.e. random effects with respect to distinct individuals, but it is worth noting that nested levels of grouping can also be considered if desired.

The random effects, serial correlation and measurement error are incorporated additively in the model equation and occur linearly in the model function. There principally are two types of random effects that are intuitively understandable: A random intercept represents the idea of each individual having its own particular level. Therefore, all measurements of the respective individual are increased or decreased by a certain quantity relative to the population average. Additionally, random slopes can be incorporated in order to allow a specific amount of growth or decline for each individual.

As both the variation between subjects and within subjects can be included in a mixed model, this kind of model is sometimes also referred to as 'two-level model of random variation'. The general definition of a linear mixed-effects model can be found in Laird and Ware (1982) and is cited and further explained by many others, among them Pinheiro and Bates (2004), Verbeke and Molenberghs (1997), Verbeke and Molenberghs (2000), Frees (2004) or Fitzmaurice *et al.* (2009). Since their first occurrence, linear mixed-effects models have been extended to incorporate various possible features of longitudinal data, such as the ones mentioned above.

For each unit $i = 1, ..., I$ there are J_i measurements at the timepoints t_{ij} for $j = 1, ..., J_i$. The used model has the general form

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{D}_i + \boldsymbol{\epsilon}_i,$$

with fixed effects $\boldsymbol{\beta}$, random effects $\boldsymbol{b}_i$ and residuals $\boldsymbol{\epsilon}_i$, as well as the matrices of covariates for the fixed and random effects, denoted by $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, respectively. Note that in general, the columns of $\boldsymbol{Z}_i$ are a subset of the columns of $\boldsymbol{X}_i$. The vector $\boldsymbol{D}_i$ covers the serial correlation structure and is seen as an independent realization from a stationary Gaussian process with mean zero, variance ξ^2 and correlation function $\rho(|t_{ij} - t_{ik}|)$, incorporating the distance between the time points of measurements j and k for the same individual i.

The random effects $\boldsymbol{b}_i$ and the residuals $\boldsymbol{\epsilon}_i$ are mutually independent Gaussian random variables, with

$$\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{J_i}) \qquad \text{and} \qquad \boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{Q}).$$

If we regard a model without serial correlation and condition on $\boldsymbol{b}_i$ and $\boldsymbol{\beta}$, the responses of one unit or individual i are independent, for which reason this model is also called 'conditional-independence model'.
Conditioning on the random subject-specific terms $\boldsymbol{b}_i$ leads to

$$\mathrm{E}(\boldsymbol{y}_i \mid \boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$$

with covariance matrix

$$\mathrm{Cov}(\boldsymbol{y}_i \mid \boldsymbol{b}_i) = \sigma^2 \boldsymbol{I}_{J_i} + \xi^2 \boldsymbol{\rho}(|\boldsymbol{u}_i|),$$

where $\boldsymbol{u}_i$ denotes the matrix containing the time distances between the measurements of the respective individual i. The mean of the marginal distribution of $\boldsymbol{y}_i$ can be obtained by calculating

$$\begin{aligned} \mathrm{E}(\boldsymbol{y}_i) &= \mathrm{E}(\mathrm{E}(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) \\ &= \mathrm{E}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{D}_i) \\ &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\,\mathrm{E}(\boldsymbol{b}_i) + \mathrm{E}(\boldsymbol{D}_i) \\ &= \boldsymbol{X}_i\boldsymbol{\beta}. \end{aligned} \tag{1}$$

Analogously, the marginal covariance matrix is derived via

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{y}_i) &= \mathrm{E}(\mathrm{Cov}(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) + \mathrm{Cov}(\mathrm{E}(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) \\ &= \mathrm{E}(\sigma^2 \boldsymbol{I}_{J_i} + \xi^2 \boldsymbol{\rho}(|\boldsymbol{u}_i|)) + \mathrm{Cov}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i). \end{aligned}$$

Using

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i) &= \mathrm{Cov}(\boldsymbol{Z}_i\boldsymbol{b}_i) \\ &= \boldsymbol{Z}_i\,\mathrm{Cov}(\boldsymbol{b}_i)\boldsymbol{Z}_i^T \\ &= \boldsymbol{Z}_i\boldsymbol{Q}\boldsymbol{Z}_i^T, \end{aligned}$$

we finally get

$$\mathrm{Cov}(\boldsymbol{y}_i) = \sigma^2 \boldsymbol{I}_{J_i} + \xi^2 \boldsymbol{\rho}(|\boldsymbol{u}_i|) + \boldsymbol{Z}_i\boldsymbol{Q}\boldsymbol{Z}_i^T. \tag{2}$$

Note that the notation can be made more compact by summarising the observations of all individuals. To do this, the subject-specific vectors and matrices have to be modified to:

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_I \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_I \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_I \end{pmatrix},$$

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \\ \vdots \\ \boldsymbol{D}_I \end{pmatrix}, \quad \boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_I \end{pmatrix} \text{ and } \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_I \end{pmatrix},$$

as well as

$$\boldsymbol{Z} = \text{diag}(\boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_I).$$

This leads to the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{D} + \boldsymbol{\epsilon}$$

with

$$\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N) \qquad \text{and} \qquad \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{W}),$$

where N represents the total number of observations in the model. The covariance matrices are blockdiagonal, and $\boldsymbol{W} = \text{diag}(\boldsymbol{Q}, ..., \boldsymbol{Q})$, where $\boldsymbol{Q}$ is repeated I times, i.e. once for each individual.

Accordingly, the summarised marginal distribution of $\boldsymbol{y}$ is then

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_N + \boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}^T + \xi^2 \boldsymbol{\rho}(|\boldsymbol{u}|)),$$

as derived above. The choices for potential serial correlation structures $\rho(|t_{ij} - t_{ik}|)$ are manifold, and the concrete decision depends strongly on the respective data set. Pinheiro and Bates (2004, p. 232) present some popular choices for continuous data. Note that these correlation structures were originally used to model spatial dependences and can also be applied with other metrics than the absolute distance, like e.g. the Euclidean norm. Among the most widely used of these alternatives are the exponential correlation model

$$\rho(|t_{ij} - t_{ik}|) = \exp(-\phi|t_{ij} - t_{ik}|),$$

or the Gaussian correlation model

$$\rho(|t_{ij} - t_{ik}|) = \exp(-\phi|t_{ij} - t_{ik}|^2)$$

for some value of $\phi > 0$. Other correlation structures have been proposed and examined with respect to their predictive abilities, e.g. by Taylor and Law (1998).

It is worth mentioning that in a model with random intercept and slope, special attention must be given to the way of coding the time variable, especially the definition of timepoint 0. Essentially, each timepoint can be declared to represent 'zero time', but this may have considerable impact on the interpretation and estimation of the random intercept. The intercept parameter describes features of the individual behaviour at the respective 'zero time'. This can for example be chosen as the time of a baseline measurement or some other kind of starting point. Alternatively, the time can be centered around a midpoint, so that the intercept reflects individual properties around the middle of the time axis. A less common but also possible choice of $t = 0$ is the last timepoint of the data set or even a time which is not contained in the range of the actual time variable. In this case the intercept parameter would represent an extrapolation of the time data.

While the differences in interpretation are relatively easy to deal with, the impact on the estimation of the respective variance is essential. Depending on the degree of individual heterogeneity at different timepoints, the variation of the random intercept for two different definitions of $t = 0$ can differ quite strongly, which is e.g. described further in Hedeker and Gibbons (2006, p. 59).

1.2 Estimation

Maximum likelihood inference for linear mixed-effects models is based on the marginal distribution for the response $\boldsymbol{y}$. In order to make the notation shorter, we denote the serial correlation function with just $\boldsymbol{\rho}$, dropping the measure of distance, and the marginal covariance matrix of $\boldsymbol{y}$ is $\boldsymbol{V} := \sigma^2 \boldsymbol{I}_N + \boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}^T + \xi^2\boldsymbol{\rho}$. Except for additive constants, the log-likelihood for the parameters $\boldsymbol{\beta}$, σ^2, $\boldsymbol{W}$, ξ^2 and $\boldsymbol{\rho}$ is

$$l(\boldsymbol{\beta}, \sigma^2, \boldsymbol{W}, \xi^2, \boldsymbol{\rho}) = -\frac{1}{2}\{\log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\}.$$

Maximising the log-likelihood with regard to $\boldsymbol{\beta}$ for fixed variance parameters σ^2, $\boldsymbol{W}$, ξ^2 and $\boldsymbol{\rho}$ leads to the estimator

$$\tilde{\boldsymbol{\beta}}(\sigma^2, \boldsymbol{W}, \xi^2, \boldsymbol{\rho}) = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y},$$

which is a weighted least-squares estimator and depends on the variance parameters. In the next step, using this estimate for $\boldsymbol{\beta}$ in $l(\boldsymbol{\beta}, \sigma^2, \boldsymbol{W}, \xi^2, \boldsymbol{\rho})$ leads to the profile log-likelihood

$$l_p(\sigma^2, \boldsymbol{W}, \xi^2, \boldsymbol{\rho}) = -\frac{1}{2}\{\log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\sigma^2, \boldsymbol{W}, \xi^2, \boldsymbol{\rho}))^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\sigma^2, \boldsymbol{W}, \xi^2, \boldsymbol{\rho}))\}.$$

Maximising the profile log-likelihood with respect to the variance parameters results in the maximum likelihood estimate.

To deduce the estimator for the individual random effects, we take the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{b}$,

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{b}\end{bmatrix} \sim N\left\{\begin{bmatrix}\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{V} & \boldsymbol{Z}\boldsymbol{W}\\ \boldsymbol{W}\boldsymbol{Z}^T & \boldsymbol{W}\end{bmatrix}\right\},$$

from which the conditional expected value

$$\mathrm{E}(\boldsymbol{b} \,|\, \boldsymbol{y}, \boldsymbol{\beta}) = \boldsymbol{W}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

can be obtained, using the properties of the multivariate normal distribution. Plugging in the estimated fixed effects $\tilde{\boldsymbol{\beta}}$ leads to the estimator

$$\mathrm{E}(\boldsymbol{b} \,|\, \boldsymbol{y}, \tilde{\boldsymbol{\beta}}) = \boldsymbol{W}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}).$$

This and alternative approaches leading to the same estimator are e.g. explained in Fahrmeir *et al.* (2007).

As pointed out regularly in the literature, among many others in Frees (2004, p. 101) or Searle *et al.* (1992, p. 249), a serious problem occurs when the maximum likelihood method for estimating variance components is used: It often underestimates the variance components and introduces a considerable bias to the estimates. The reason for this is that the loss of degrees of freedom that the estimation of the fixed effects induces is not considered by the maximum likelihood method. Therefore it is relatively seldom used, the more so as there is an alternative estimating method which omits this bias: the restricted maximum likelihood method.

The restricted maximum likelihood (REML) estimator is obtained via the marginal or restricted log-likelihood and can be transformed so that it adds an additional parameter to the profile log-likelihood:

$$l_r(\sigma^2, \boldsymbol{W}, \xi^2, \boldsymbol{\rho}) = \log\left(\int L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{W}, \xi^2, \boldsymbol{\rho}) d\boldsymbol{\beta} \right)$$
$$= l_p(\sigma^2, \boldsymbol{W}, \xi^2, \boldsymbol{\rho}) - \frac{1}{2} \log |\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}|. \quad (3)$$

Despite the obvious benefits of REML estimation, it has to be applied with caution and depending on the purpose of fitting the respective model, because the comparison of linear mixed-effects models with different fixed-effects structures is impossible if REML was used to fit these models. The non-restricted likelihood function is invariant to reparameterisations of the fixed effects, whereas the restricted likelihood function changes when the design matrix $\boldsymbol{X}$ is altered.

2 Generalized linear mixed models for longitudinal data

The concept of linear mixed models presented in Section 1 can easily be extended to discrete observations, leading to generalized linear mixed models (GLMMs). As the inclusion of serial correlation is by far less common in the case of generalized mixed models, we omit the serial correlation part in the following subsections, however, its inclusion in the models is straightforward.

2.1 Model formulation

The basic definition of a generalized linear mixed model works as follows (for more information see e.g. Fahrmeir and Tutz, 2001, Molenberghs and Verbeke, 2005, or Fitzmaurice *et al.*, 2009):

For each individual $i = 1, \ldots, I$, discrete observations y_{ij} at time points t_j with $j = 1, \ldots, J$ are available. For the following explications, we assume that each individual provides the same number of observations at the same time points, but this is not generally necessary. The vectors $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ contain covariates relating to the fixed and random effects, respectively, and the linear predictor is defined as

$$\eta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i, \quad (4)$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the fixed effects and $\boldsymbol{b}_i \in \mathbb{R}^q$ represent the random effects. The latter are assumed to be normally distributed $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{Q})$ and that different measurements of the same subject are independent, conditional on the random effects $\boldsymbol{b}_i$. Their conditional expected value $\mu_{ij} = \mathrm{E}(y_{ij} \mid \boldsymbol{b}_i)$ is related to the linear predictor η_{ij} via an appropriate link function g, so that

$$g(\mu_{ij}) = \eta_{ij}. \quad (5)$$

Given both fixed and random effects, the conditional distribution of the response y_{ij} belongs to an exponential family with density

$$f(y_{ij} \mid \boldsymbol{b}_i) = \exp\left\{ \frac{y_{ij}\theta_{ij} - \kappa(\theta_{ij})}{\phi} + c(y_{ij}, \phi) \right\},$$

with natural parameter θ_{ij}, dispersion parameter ϕ and two functions κ and c that depend on the type of exponential family. The conditional expected value μ_{ij} equals $\kappa'(\theta_{ij})$, where κ' denotes the first derivative of κ and depends on $\boldsymbol{b}_i$ via (4) and (5).

Two non-Gaussian regression models that occur quite often in the context of generalized linear mixed models are the binary logistic and log-linear Poisson regression models: For binary data, it is assumed that each observation y_{ij} has a Bernoulli distribution with probability

$$p_{ij} = P\{y_{ij} = 1\} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})},$$

leading to a binary logistic regression model. In the case of a log-linear Poisson model for count data, the expectation

$$\lambda_{ij} = \exp(\eta_{ij}).$$

is assumed.

Note that the assumed normal distribution of the random effects $\boldsymbol{b}_i$ is hard to verify and might be wrong in many cases. Fitzmaurice *et al.* (2009, p. 339) explain that the estimates of the random effects are naturally sensitive to the specification of their distribution. However, the fixed effects are barely affected by wrong distributional assumptions of the random effects, so that the estimated effects on population level are reliable enough. Details on inference in generalized linear mixed models can be found in the following subsection.

2.2 Estimation

As in linear mixed models, maximum likelihood estimation is the principal tool for inference in generalized linear mixed models. In contrast to the linear case, however, maximizing the likelihood cannot be done analytically in most cases because the expected value μ_{ij} depends non-linearly on the linear predictor η_{ij}. Therefore, numerical methods have to be applied, which can be quite involved and time consuming, depending on the actual model and the size of the data set.

The principal idea of maximum likelihood estimation in generalized linear mixed models is analogous to the case of linear mixed models. Estimates of the fixed effects $\boldsymbol{\beta}$ and the covariance matrix of the random effects $\boldsymbol{Q}$ are obtained by maximizing the marginal likelihood function

$$L(\boldsymbol{\beta}, \boldsymbol{Q}) = \prod_{i=1}^{I} \int \prod_{j=1}^{J} f(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\beta}) f(\boldsymbol{b}_i | \boldsymbol{Q}) d\boldsymbol{b}_i \tag{6}$$

or the marginal log-likelihood

$$l(\boldsymbol{\beta}, \boldsymbol{Q}) = \sum_{i=1}^{I} \log \int \prod_{j=1}^{J} f(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\beta}) f(\boldsymbol{b}_i | \boldsymbol{Q}) d\boldsymbol{b}_i. \tag{7}$$

To do this, estimates of the random effects $\boldsymbol{b}_i$ have to be provided as well by calculating the posterior conditional modes

$$\hat{\boldsymbol{b}}_i = \arg\max_{\boldsymbol{b}_i} f(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \hat{\boldsymbol{\beta}}) f(\boldsymbol{b}_i | \hat{\boldsymbol{Q}})$$

for given estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{Q}}$, as explained in Bates (2013). This means that estimates of $\boldsymbol{\beta}$, $\boldsymbol{Q}$ and $\boldsymbol{b}_i$ are calculated alternately until convergence. Several numerical methods have been proposed for this task. A useful overview of different approaches to maximizing the marginal likelihood as for example Gauss-Hermite approximation or Monte Carlo integration is given in Molenberghs and Verbeke (2005, p. 268) and Fahrmeir and Tutz (2001, chapter 7).

Gauss-Hermite approximation is one of the most commonly used methods and is applied in all examples in this thesis. Its basic idea is to approximate the integral in equation (6) with a weighted sum, using a number of quadrature points and their associated weights. The adaptive Gauss-Hermite quadrature is based on the marginal log-likelihood (7) and requires fewer quadrature points. The number of quadrature points is crucial for the accuracy of the approximations, and as always a tradeoff between reliability and computing time has to be found.

In the `lme4` package, the calculation of the posterior conditional modes in equation (2.2) is done using a penalized iteratively reweighted least squares (PIRLS) algorithm (Bates and DebRoy, 2004). This is the same as the so-called iteratively weighted least squares algorithm with prior distribution proposed in Gamerman (1997) and works as follows:

For the estimation of the random effects $\boldsymbol{b}_i$ of individual i, we treat $\boldsymbol{x}_i\hat{\boldsymbol{\beta}}$ as an offset and use the matrix $\boldsymbol{z}_i$ of dimension $q \times J$ as design matrix. Combining the prior distribution $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \hat{\boldsymbol{Q}})$ with the likelihood leads to the posterior distribution

$$\boldsymbol{b}_i \mid \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{Q}} \overset{a}{\sim} N(\boldsymbol{m}_i, \boldsymbol{C}_i).$$

After choosing some starting values $\boldsymbol{m}^{(0)}$, the estimates $\boldsymbol{m}_i^{(k)}$ and $\boldsymbol{C}_i^{(k)}$ in iteration step k are

$$\boldsymbol{m}_i^{(k)} = \boldsymbol{C}_i^{(k)} \boldsymbol{z}_i \boldsymbol{W}_i(\boldsymbol{m}_i^{(k-1)}) \tilde{\boldsymbol{y}}_i(\boldsymbol{m}_i^{(k-1)})$$

and

$$\boldsymbol{C}_i^{(k)} = \{\hat{\boldsymbol{Q}}^{-1} + \boldsymbol{z}_i \boldsymbol{W}_i(\boldsymbol{m}_i^{(k-1)}) \boldsymbol{z}_i^T\}^{-1}$$

with response vector $\tilde{\boldsymbol{y}}_i(\boldsymbol{m}_i^{(k-1)})$ that contains pseudo observations

$$\tilde{y}_{ij}(\boldsymbol{m}_i^{(k-1)}) = \boldsymbol{z}_{ij}^T \boldsymbol{m}_i^{(k-1)} + \{y_{ij} - \mu_{ij}(\boldsymbol{m}_i^{(k-1)})\} g'\{\mu_{ij}(\boldsymbol{m}_i^{(k-1)})\},$$

where $\mu_{ij}(\boldsymbol{m}_i^{(k-1)}) = g^{-1}(\boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T \boldsymbol{m}_i^{(k-1)})$. The components of the weight matrix $\boldsymbol{W}_i(\boldsymbol{m}_i^{(k-1)}) = \text{diag}\{W_{ij}(\boldsymbol{m}_i^{(k-1)})\}$ are defined via

$$W_{ij}^{-1}(\boldsymbol{m}_i^{(k-1)}) = \kappa''(\theta_{ij}(\boldsymbol{m}_i^{(k-1)}))\{g'(\mu_{ij}(\boldsymbol{m}_i^{(k-1)}))\}^2.$$

The iterations should be stopped once some predefined convergence criterion is fulfilled. The estimates $\hat{\boldsymbol{m}}_i$ are used as estimated random effects $\hat{\boldsymbol{b}}_i$.

Concerning restricted maximum likelihood estimation, it can be shown that the restricted log-likelihood for generalized linear mixed models looks similar to the linear case (3), but working observations

$$\tilde{y}_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i + d_{ij}^{-1}(y_{ij} - \mu_{ij})$$

with

$$d_{ij} = \frac{dg^{-1}(\mu_{ij})}{d\eta_{ij}}$$

are used instead of the actual observations y_{ij} (see Kneib, 2006, p. 71).

3 Model choice in linear and generalized linear mixed models

The log-likelihood which was presented in the former section is not only necessary for estimation purposes, but also concerning the issue of model choice. However, the log-likelihood itself, here generally denoted by $\log L$ is not suited for model choice due to the fact that it increases automatically with higher model complexity. Instead, it is incorporated in the two most frequently used criteria for model choice, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Let N be the number of observations in a data set and p the number of parameters in the model. Then the form of the AIC and BIC is

$$\text{AIC} = -2\log L + 2p$$

and

$$\text{BIC} = -2\log L + p\log(N).$$

In this formulation the two criteria are negatively oriented, so that the rule for model selection is to choose the model with the lowest AIC or BIC, respectively. The two criteria look very similar on first sight, but their theoretical background is quite different. The AIC was suggested by Akaike (1973) and is based on the Kullback-Leiber distance between two probability distributions. It is linked to the idea of cross-validation, as it can be shown that it is (up to a multiplicative factor) asymptotically equivalent to the mean value of the cross-validated mean logarithmic score (for more information see e.g. Stone, 1977 or Pawitan, 2001, p. 381; the logarithmic score is presented in Section 4 of this introduction).

By contrast, the Bayesian information criterion, which is also known as Schwarz criterion, is derived (as its name suggests) from a purely Bayesian point of view and closely related to the marginal likelihood as well as to Bayesian model choice based on so-called Bayes factors (see e.g. also Kass and Raftery, 1995). Generally, the AIC selects models with more parameters than the BIC, because model complexity is punished more severly by the latter criterion.

Unfortunately, model choice in linear and generalized linear mixed models based on classical criteria like AIC or BIC is not as straightforward as it is in models without random effects. In the latter case the goal is to simply select the most relevant covariates out of a list of potentially influential variables, but there are more challenges when a mixed model is concerned: Apart from the choice of covariates that are included in the fixed effects, the random effects that shall be included in the model also have to be selected. An overview of the issues associated with model choice in mixed models is given in Claeskens and Hjort (2008, Section 10.1).

The presence of these two tasks greatly influences the concepts of model choice in general, and the actual construction of the AIC and the BIC in particular. Depending on the main focus of the desired analysis, there are two distinct likelihoods that can be chosen for use in the respective criterion for model choice, namely the marginal and the conditional likelihood. The marginal likelihood should be chosen when the main interest lies in inference on the population, because it refers to the average behavior of the whole population. Individual random effects are integrated out and can therefore not be estimated and taken into account when predictions are made. By contrast, the primary focus of the conditional likelihood is on individual random effects, so that individual estimates and predictions are possible.

Depending on the appropriate likelihood for the respective question, the appearance of the AIC and BIC essentially changes: the general definitions of these model choice criteria have to be adapted when their use for mixed-effects models is desired. First of all, the associated marginal or conditional likelihood has to be applied, depending on the context and the concrete problem. While in the case of inference for the population mean, we can use the

marginal likelihood and add the number of fixed parameters in the penalty term, in the case where estimates of the individual random effects are desired or questions on the inclusion of a particular random effect shall be answered, the conditional likelihood has to be applied and the penalty term needs to be adapted.

Several authors have worked on the applicability of the AIC and BIC for linear mixed models: Vaida and Blanchard (2005) explain the problems associated with the application of the conditional AIC (cAIC) in more detail and present its calculation and properties for the case where the (scaled) covariance matrix of the random effects is known. Hodges and Sargent (2001) show how the effective degrees of freedom which determine the penalty term of the cAIC can be obtained. The effective degrees of freedom correspond to the trace of the hat matrix and represent the effective number of parameters in a linear mixed model. A general formula for the cAIC for which the variance components do not have to be known explicitly, because the uncertainty induced by estimating them can be taken into account is provided by Liang *et al.* (2008). This way, the results are much more reliable, however, this gain comes at a cost, because the penalty term needs to be calculated numerically. The resulting computational burden is immense, especially for large data sets, as shown by Greven and Kneib (2010). This is especially problematic when several candidate models are to be compared.

Based on the marginal likelihood, Pauler (1998) suggests a practical solution to this problem by using a modified version of the BIC, where the special structure of unbalanced logitudinal data is incorporated in the penalty term. Therefore, $p \log(N)$ is replaced by a sum whose summands depend on whether a covariate has an associated random effect or not:

$$\mathrm{BIC}_{\mathrm{mod}} = -2 \log L + \sum_{p=1}^{P} N_p.$$

In this formula, P represents the number of fixed effect parameters including the intercept. If the respective parameter has an associated random effect, N_p equals the number of individuals, whereas the number of observations has to be used for N_p if there is no random effect for this parameter.

However, this allows only the choice of the fixed effects parameters, but no decision on the necessity of a random effect can be obtained this way, because replacing the marginal by the conditional likelihood makes the correct adaptation of the penalty term impossible. Pauler *et al.* (1999) present another approximation to the BIC that can be used for selecting random effects. An additional boundary correction term is introduced, using a boundary Laplace approximation. A generalized information criterion for model selection in linear mixed models is suggested by Pu and Niu (2006), however, the associated penalty term involves only the complete number of observations without accounting for the number of fixed and random effects parameters. Moreover, simulations in their article show that this criterion is primarily useful for the choice of fixed effects only.

While these are potentially useful modifications of the classical model choice criteria for the choice of covariates and random effects, to our knowledge, no such modification has been developed concerning the choice of additional serial correlation structure. It is generally recommended to choose the correlation structure based on the empirical variogram of suitable residuals (see e.g. Diggle *et al.*, 2002, p. 85) or by assessing the model fit (see Verbeke and Molenberghs, 2000, p. 137). However, none of these methods provide a formal model choice criterion.

When it comes to the applicability of the AIC and BIC to generalized linear mixed models, the literature is more sparse. The main problem in this context is that there is no analytic

deduction of the conditional AIC, so that one has to resort to asymptotic measures. Two very similar approximations of the cAIC are suggested by Donohue *et al.* (2011) and Yu and Yau (2012), but these approximations have some disadvantages. Apart from the fact that the asymptotics might be problematic for small data sets, the calculation of these criteria involves multiplication and inversion of large matrices, which is time-consuming or even impossible if a large data set is involved. The size of the data set is also a limiting factor in the calculation of an unbiased estimator of cAIC for Poisson regression models proposed by Lian (2012), because a considerable number of model fits is required. The use of fence methods, which restrict the number of candidate models by applying increasingly strict boundaries on some criterion, is suggested by Jiang *et al.* (2008) and Nguyen and Jiang (2012).

4 Proper scoring rules

The present section aims to introduce several measures that can be used to assess the predictive quality of all statistical models, and of longitudinal mixed-effects models in particular. These measures are not only useful for gaining information on predictive abilities in general, but they can also be a potential solution to concrete problems such as whether a special type of random effect should be included in a linear mixed-effects model or not.

An ideal predictive distribution should fulfill two main tasks: On the one hand the predictive distribution should be as sharp as possible, while on the other hand it should be well calibrated. Sharpness is a property of the predictions: It refers to the concentration of the predictive distribution. The sharper a distribution, the smaller is the range of possible predicted values. A distribution that lacks sharpness generally is of little use, because the predictions are too imprecise and therefore not informative.

On the other hand, calibration is a joint property of the predictive distribution and the real data and stands for their agreement. Ideally, the true values that materialize later are consistent with the chosen predictive distribution. A predictive distribution can indeed be very sharp, but despite that it is of no use, if the distribution fails to represent the true values because it is skew or misplaced. Therefore, as Gneiting and Raftery (2007) put it, a desirable predictive distribution should be as sharp as possible, subject to calibration. In the following subsections, tools that can be used to control for sharpness and calibration and are useful for comparing distinct forecasting strategies are presented.

The essential quantities to compare a predictive distribution Y with the truely observed value y_{obs} are proper scores which are also called proper scoring rules. They allow to calculate a numerical value for each of the competing models which can then be compared in order to determine the model with the best fit. They are typically positively oriented, so that a larger value denotes the better model and can be seen as a kind of reward for assuming a realistic model. Gneiting *et al.* (2007) point out that scoring rules cover both sharpness and calibration at once. Here, $S(Y, y_{\text{obs}})$ stands for a score related to Y and the true value y_{obs}, while $\mathrm{E}(S(Y, U))$ denotes the expected value of $S(Y, y_{\text{obs}})$, given the true predictive distribution U.

A very important term in the context of scoring rules is propriety, which is defined as follows: The expected value of a proper score becomes maximal if the observed value is in fact derived from the assumed distribution and not any other one. A score is even strictly proper if this maximum is unique. This fact can also be expressed more formally: Let U be the true and best predictive distribution and U' any other possible predictive distribution. Then propriety is denoted by

$$S(U', U) \leq S(U, U) \text{ for all } U, U',$$

and

$$S(U', U) = S(U, U)$$

if and only if $U = U'$. This definition also has a concrete practical interpretation, which can be found in Garthwaite *et al.* (2005) among others: Proper scores do not lead the forecaster to turn away from his true belief. If a score is strictly proper, such a shift of opinion would even be penalized.

Bröcker and Smith (2007) name the reason why proper scores should be used in practice by pointing out that only proper scores always prefer predictive distributions that are actually better suited for the concrete forecasting problem. They explain that improper scores can possibly act in ways that are not anticipated and cannot be understood with common sense. A convenient characteristic of proper scoring rules is the fact that the mean of proper scores remains proper, which allows to summarize the scores of the predictions for several measurements. Therefore, competing models can be compared quickly using just one mean value.

Perhaps the best known and most frequently used proper score is the Brier score (Brier, 1950) for binary predictions. It can be applied for binary data and can alternatively be called quadratic score. It is defined as

$$\mathrm{BS}(Y, y_{\mathrm{obs}}) = -(p_Y - y_{\mathrm{obs}})^2, \tag{8}$$

where p_Y stands for the predicted probabiliy of the outcome and y_{obs} is the actual binary observation. Its interpretation as the squared distance between a probability and the actual binary observation is easy to understand and quite intuitive. This and the easy calculation are the main reasons why this score is so popular.

To prove that the Brier score is strictly proper, we calculate the Brier score for the true predictive distribution U and some other predictive distribution U'. Its expected value is then

$$\begin{aligned} \mathrm{E}(\mathrm{BS}(U', U)) &= -p_U(p_{U'} - 1)^2 - (1 - p_u)p_{U'}^2 \\ &= 2p_{U'}p_U - p_U - p_{U'}^2. \end{aligned}$$

Differentiating with respect to $p_{U'}$ leads to

$$\frac{d\,\mathrm{E}(\mathrm{BS}(U', U))}{dp_{U'}} = 2p_U - 2p_{U'}.$$

This expression equals zero only if $p_U = p_{U'}$ The second derivative is

$$\frac{d^2\,\mathrm{E}(\mathrm{BS}(U', U))}{dp_{U'}^2} = -2,$$

from which we can see that the maximum is unique.

Another common proper score is the logarithmic score (denoted here by LS) which simply evaluates the logarithmic predictive density of the effectively observed value y_{obs}:

$$\mathrm{LS}(Y, y_{\mathrm{obs}}) = \log f_Y(y_{\mathrm{obs}}) \tag{9}$$

The logarithmic score is sometimes also called ignorance score due to its possible interpretation as measure for the average information deficit. As explained from an information theoretic point of view in Roulston and Smith (2002), it is closely linked to the Shannon entropy and can therefore be seen as measure for the distance between two predictive distributions,

the true one and the one that is actually used, with respect to the information they contain.

The logarithmic score is one of the scores that are frequently used for continuous data, because it is quickly and easily computed once the predictive distribution is explicitly specified. That the logarithmic score is strictly proper can be proved by looking at the difference of the expected values for two predictive distributions U and U', i.e. $\mathrm{E}(\mathrm{LS}(U,U)) - \mathrm{E}(\mathrm{LS}(U',U))$. Using the explicit expressions for the expected values, this difference becomes

$$\begin{aligned}\mathrm{E}(\mathrm{LS}(U,U)) - \mathrm{E}(\mathrm{LS}(U',U)) &= \int_{-\infty}^{\infty} \log\big(f_U(y)\big) f_U(y) dy - \int_{-\infty}^{\infty} \log\big(f_{U'}(y)\big) f_U(y) dy \\ &= \int_{-\infty}^{\infty} \Big(\log\big(f_U(y)\big) - \log\big(f_{U'}(y)\big)\Big) f_U(y) dy \\ &= \int_{-\infty}^{\infty} f_U(y) \log\left(\frac{f_U(y)}{f_{U'}(y)}\right) dy \geq 0,\end{aligned}$$

with equality if and only if $U = U'$. This holds due to the information inequality (Kullback and Leibler, 1951, Lemma 3.1). As the predictive distribution is only evaluated at the actual observation y_{obs}, the LS is a local score, which ignores all other values of $f_Y(y)$. For this reason its concept is regularly criticized as being too limited. Moreover, the logarithmic score is very sensitive to outlying observations, which is not always desired.

Another proper scoring rule for univariate predictive distributions is the so-called continuous ranked probability score (CRPS). It is the integral over all possible thresholds r of the Brier Score (8) and takes the form

$$\mathrm{CRPS}(Y, y_{\text{obs}}) = -\int_{-\infty}^{\infty} (P_Y(Y \leq r) - \mathbf{1}(y_{\text{obs}} \leq r))^2 dr.$$

In contrast to the LS, this score is sensitive to distance, which means that it takes into account how close a predicted value is to the observed value, rewarding models which assign high probabilities to values next to the true value (see also Gneiting and Raftery, 2007 and Hersbach, 2000).

The CRPS is strictly proper for probability measures with finite first moment, which can be seen following Gneiting and Raftery (2007): Let $x \in \mathbb{R}$ and $Y, Y' \overset{iid}{\sim} F$ with existing and finite first moment, from which follows the existence of a finite absolute first moment. Then

$$\begin{aligned}\mathrm{CRPS}(F,x) &= -\int_{-\infty}^{\infty} \Big(F(r) - \mathbf{1}(x \leq r)\Big)^2 dr \\ &= \frac{1}{2}\,\mathrm{E}\,|Y - Y'| - \mathrm{E}\,|Y - x|.\end{aligned}$$

The expected values of the CRPS have to be calculated for two different situations. If the observation comes from the correct predictive distribution F, i.e. $X, X', Y, Y' \overset{iid}{\sim} F$, the expected value has the form

$$\begin{aligned}\mathrm{E}(\mathrm{CRPS}(F,F)) &= \frac{1}{2}\,\mathrm{E}\,|Y - Y'| - \mathrm{E}\,|Y - X| = \frac{1}{2}\,\mathrm{E}\,|Y - Y'| - \mathrm{E}\,|Y - Y'| \\ &= \frac{1}{2}\,\mathrm{E}\,|X - X'| - \mathrm{E}\,|X - X'| = -\frac{1}{2}\,\mathrm{E}\,|X - X'|.\end{aligned}$$

If, on the other hand, the predictive distribution is different from the observation's true distri-

bution, i.e. $Y, Y' \overset{iid}{\sim} G$ and $X \sim F$, we get

$$\mathrm{E}(\mathrm{CRPS}(G, F)) = \frac{1}{2}\,\mathrm{E}\,|Y - Y'| - \mathrm{E}\,|Y - X|.$$

Taking the difference of these two expressions leads to the divergence function

$$\mathrm{E}(\mathrm{CRPS}(F, F)) - \mathrm{E}(\mathrm{CRPS}(G, F)) = -\frac{1}{2}\,\mathrm{E}\,|X - X'| - \frac{1}{2}\,\mathrm{E}\,|Y - Y'| + \mathrm{E}\,|Y - X|.$$

Using Baringhaus and Franz (2004, Lemmas 2.1 and 2.2), $E|Y - X|$ can also be written as

$$\mathrm{E}\,|Y - X| = \int G(u)(1 - F(u)) + F(u)(1 - G(u))du,$$

and analogously

$$\mathrm{E}\,|X - X'| = \int 2F(u)(1 - F(u))du \quad \text{and} \quad \mathrm{E}\,|Y - Y'| = \int 2G(u)(1 - G(u))du.$$

This leads to

$$\begin{aligned}
&\mathrm{E}(\mathrm{CRPS}(F, F)) - \mathrm{E}(\mathrm{CRPS}(G, F)) \\
&\quad = \int G(u)(1 - F(u)) + F(u)(1 - G(u)) - F(u)(1 - F(u)) - G(u)(1 - G(u))du \\
&\quad = \int G(u) - F(u)G(u) + F(u) - F(u)G(u) - F(u) + F(u)^2 - G(u) + G(u)^2 du \\
&\quad = \int F(u)^2 - 2F(u)G(u) + G(u)^2 du \\
&\quad = \int (F(u) - G(u))^2 du.
\end{aligned}$$

It is obvious that this divergence function is always non-negative and can only be equal to 0 if $F(u) = G(u)$. Therefore, the CRPS is strictly proper for probability measures with finite first moment.

The CRPS has not been used frequently because the integral is often difficult to obtain. However, Gneiting and Raftery (2007) showed, using results from Baringhaus and Franz (2004) or, alternatively, Székely and Rizzo (2005), that the univariate CRPS can also be written as

$$\mathrm{CRPS}(Y, y_{\mathrm{obs}}) = \frac{1}{2}E|Y - Y'| - E|Y - y_{\mathrm{obs}}|, \tag{10}$$

where Y and Y' are independent realisations from the respective predictive density. Note that this result is only valid in the univariate case and cannot be generalized to multivariate distributions.

If the predictive distribution is univariate Gaussian, an explicit and very convenient form of the CRPS can be obtained, based on the CRPS of the form given in (10). In that case, the differences $Y - Y'$ and $Y - y_{\mathrm{obs}}$ are also normally distributed

$$Y - Y' \sim N(0, 2\sigma^2) \quad \text{and} \quad Y - y_{\mathrm{obs}} \sim N(\mu - y_{\mathrm{obs}}, \sigma^2).$$

Accordingly, the absolute value of these expressions follows a folded normal distribution with the same parameters:

$$|Y - Y'| \sim FN(0, 2\sigma^2) \qquad \text{and} \qquad |Y - y_{\text{obs}}| \sim FN(\mu - y_{\text{obs}}, \sigma^2).$$

The expected value of a folded normal distribution with parameters μ and σ^2 has the general form

$$\mathrm{E}(X) = 2\sigma\phi(\mu/\sigma) + \mu(2\Phi(\mu/\sigma) - 1),$$

where φ and Φ represent the density function or the cumulative distribution function of a standard normal distributed variable, respectively. This expression reduces to

$$\mathrm{E}(X) = \sigma\sqrt{2/\pi}$$

if $\mu = 0$.
Using these expressions of the expected value of a folded normal distribution, we get

$$\begin{aligned}
\mathrm{CRPS}(Y, y_{\text{obs}}) &= \frac{1}{2}E|Y - Y'| - E|Y - y_{\text{obs}}| \\
&= \frac{1}{2}\sqrt{2}\sigma\sqrt{\frac{2}{\pi}} - 2\sigma\phi\Big(\frac{\mu - y_{\text{obs}}}{\sigma}\Big) - (\mu - y_{\text{obs}})\Big(2\Phi\Big(\frac{\mu - y_{\text{obs}}}{\sigma}\Big) - 1\Big) \\
&= \frac{\sigma}{\sqrt{\pi}} - 2\sigma\phi\Big(\frac{y_{\text{obs}} - \mu}{\sigma}\Big) - (\mu - y_{\text{obs}})\Big(2\Big(1 - \Phi\Big(\frac{y_{\text{obs}} - \mu}{\sigma}\Big)\Big) - 1\Big) \\
&= \sigma\left[\frac{1}{\sqrt{\pi}} - 2\varphi\Big(\frac{y_{\text{obs}} - \mu}{\sigma}\Big) - \frac{y_{\text{obs}} - \mu}{\sigma}\left(2\Phi\Big(\frac{y_{\text{obs}} - \mu}{\sigma}\Big) - 1\right)\right].
\end{aligned}$$

Unfortunately, this way of obtaining the CRPS is no longer possible if the predictive distribution is multivariate.
The last score we use is the Dawid-Sebastiani score, which was presented by Dawid and Sebastiani (1999) and further explained by Gneiting and Raftery (2007). It has the form

$$\mathrm{DSS}(Y, y_{\text{obs}}) = -\frac{1}{2}\Big\{\log(\sigma_Y^2) + \Big(\frac{y_{\text{obs}} - \mu_Y}{\sigma_Y}\Big)^2\Big\}$$

and is equivalent to the LS (9) under normality. The DSS is strictly proper for distributions with finite first and second moments, which can be seen as follows: The expected value of the DSS is

$$\begin{aligned}
\mathrm{E}(\mathrm{DSS}(U', U)) &= -\log(\sigma_{U'}) - \frac{1}{2}\,\mathrm{E}\Big(\frac{y_{\text{obs}}^2 - 2y_{\text{obs}}\mu_{U'} + \mu_{U'}^2}{\sigma_{U'}^2}\Big) \\
&= -\log(\sigma_{U'}) - \frac{1}{2}\Big(\frac{1}{\sigma_{U'}^2}\,\mathrm{E}(y_{\text{obs}}^2) - \frac{2\mu_{U'}}{\sigma_{U'}^2}\,\mathrm{E}(y_{\text{obs}}) + \frac{\mu_{U'}}{\sigma_{U'}^2}\Big).
\end{aligned}$$

The definition of the variance $\sigma_U^2 = \mathrm{E}(y_{\text{obs}}^2) - \mu_U^2$ is used to get the expression $\mathrm{E}(y_{\text{obs}}^2)$, so that we get

$$\mathrm{E}(\mathrm{DSS}(U', U)) = -\log(\sigma_{U'}) - \frac{\sigma_U^2 + \mu_U^2 - 2\mu_U\mu_{U'} + \mu_{U'}^2}{2\sigma_{U'}^2}.$$

The first derivatives of $\mathrm{E}(\mathrm{DSS}(U', U))$ with respect to $\mu_{U'}$ and $\sigma_{U'}$ are

$$\frac{d\,\mathrm{E}(\mathrm{DSS}(U', U))}{d\mu_{U'}} = \frac{\mu_U - \mu'_U}{\sigma^2_{U'}}$$

and

$$\frac{d\,\mathrm{E}(\mathrm{DSS}(U', U))}{d\sigma_{U'}} = \frac{\sigma^2_U - \sigma^2_{U'} + \mu^2_U - 2\mu_U\mu_{U'} + \mu^2_{U'}}{\sigma^3_{U'}}.$$

Setting these derivatives equal to zero leads to $\mu_U = \mu_{U'}$ and $\sigma_U = \sigma_{U'}$. We show that this is a maximum by looking at the according Hessian matrix

$$\mathrm{H}(\mathrm{E}(\mathrm{DSS}(U', U))) = \begin{pmatrix} -\frac{1}{\sigma^2_{U'}} & 2\frac{\mu_{U'} - \mu_U}{\sigma^3_{U'}} \\ 2\frac{\mu_{U'} - \mu_U}{\sigma^3_{U'}} & \frac{1}{\sigma^2_{U'}} - \frac{3\sigma^2_U}{\sigma^4_{U'}} - \frac{3(\mu_U - \mu_{U'})^2}{\sigma^4_{U'}} \end{pmatrix}.$$

Plugging in $\mu_U = \mu_{U'}$ and $\sigma_U = \sigma_{U'}$ simplifies the Hessian to

$$\mathrm{H}(\mathrm{E}(\mathrm{DSS}(U', U))) = \begin{pmatrix} -\frac{1}{\sigma^2_U} & 0 \\ 0 & -\frac{2}{\sigma^2_U} \end{pmatrix},$$

which is negative definite, proving that there is a unique maximum.

Thesis Summary

This thesis consists of three papers. Their content and contribution are briefly summarized below.

Paper I

Predictive crossvalidation for the choice of linear mixed-effects models with application to data from the Swiss HIV Cohort Study by J. Braun, L. Held and B. Ledergerber.
This paper introduces a new approach to selecting fixed effects, random effects and a suitable serial correlation structure in linear mixed models. The proposed crossvalidation procedure is computationally efficient, involving only one model fit in total. Model choice is based on mean proper scoring rules as explained in Gneiting and Raftery (2007) and Gneiting *et al.* (2007), thus choosing the components of a linear mixed model with respect to its predictive abilities. The methodology is applied to a data set from the Swiss HIV Cohort Study to select a model for the prediction of CD4+ counts, a measure for the strength of the immune system, in HIV positive patients. The close relation of the mean logarithmic score with the so-called conditional AIC is illustrated, and comparisons with alternative criteria and with the results of a full crossvalidation approach are given.

This work is based on an idea by L. Held to perform predictive crossvalidation in a similar way as Marshall and Spiegelhalter (2003) and use proper scoring rules for model comparison. Details of the methodology were jointly developed by L. Held and me. Additionally, I deducted the bias correction for AIC and the form of the hat matrix in linear mixed models with serial correlation as given in the web-based supplementary material. I implemented the procedures, conducted all analyses and wrote a draft of the manuscript, to which L. Held contributed a justification of the crossvalidation approach. B. Ledergerber provided the data set and rel-

evant medical knowledge. Both B. Ledergerber and L. Held commented on the manuscript which I subsequently finalized.

The main contribution of this paper is the introduction of a quick crossvalidation approach which greatly facilitates predictive model choice in comparison to full crossvalidation. Besides, it addresses the problem of choosing the components of a linear mixed model from a new perspective, thus stating the importance of predictions in a medical and general context.

Paper II

Choice of generalized linear mixed models using predictive crossvalidation by J. Braun, D. Sabanés Bové and L. Held.

This paper extends the predictive crossvalidation approach to the case of generalized linear mixed models. The principles of predictive crossvalidation stay basically the same, however, the necessary cross-validated predictive distribution can in most cases not be calculated analytically any more. Therefore, we propose to use a Bayesian iteratively weighted least squares (IWLS, see Gamerman, 1997) algorithm for this task. Details on its applicability for the two most common generalized linear mixed models, i.e. binary logistic and log-linear Poisson regression are discussed, and the methodology is applied to two standard data sets known from the literature, e.g. from Diggle *et al.* (2002). A comparison with alternative methods is also given.

This work is based on the ideas presented in Paper I. L. Held discussed the idea of using an IWLS algorithm with prior distribution to obtain the leave-one-out predictive distribution with D. Sabanés Bové, who developed further details of this approach. The algorithm was implemented by me, along with the remaining functions for crossvalidation and calculating mean proper scoring rules. I conducted all calculations and drafted the manuscript. Both L. Held and D. Sabanés Bové read and commented on the manuscript, which was finalized by me.

The main contribution of this paper is the extension of the predictive crossvalidation approach to a much broader set of models. The conditional AIC which is necessary especially for the choice of random effects cannot be obtained analytically, and approximate calculations are not always possible (e.g. due to the size of the data set or the presence of overdispersion). Our proposed criterion provides an attractive alternative. Although the calculation of the leave-one-out predictive distribution is more involved in generalized linear mixed models, our methodology provides a comparably quick way of obtaining a criterion for model choice.

Paper III

Accounting for baseline differences and measurement error in the analysis of change over time by J. Braun, L. Held, B. Ledergerber and the Swiss HIV Cohort Study.

This paper deals with another well-known problem of longitudinal modelling that is particularly relevant in the clinical practice: If a comparison of change over time in several groups is desired, it should be made sure that these groups are comparable with respect to their baseline value. This involves two distinct issues: Comparisons of change should be with respect to the same baseline value in all groups, and this baseline value is usually subject to measurement error, especially in observational studies. By fitting a linear mixed model and subsequently calculating the expected change from baseline, conditional on the unobserved true baseline value, both concerns can be accounted for. A previous approach by Harrison *et al.* (2009) is adapted to cover a broader set of linear mixed models, and is developed further for bivariate

models, thus providing a solution if two (or more) variables are to be modelled jointly. To illustrate the methodology, we use it to answer the question if there are differences concerning the change of CD4+ counts in HIV positive individuals with or without a coinfection with hepatitis C.

The idea for this paper came up when discussing current issues of HIV research with B. Ledergerber and L. Held. I conducted a literature review and found the paper by Harrison *et al.* (2009). As suggested by L. Held, I extended the methodology, implemented it and conducted all analyses. A first draft written by me, using also clinically relevant information by B. Ledergerber, was commented by L. Held. Afterwards, I finalized the paper.

The main contribution of this paper is that the problem of correctly analysing change from baseline is tackled for the situation where more than one measurement post baseline is available. So far, change score and ANCOVA methods are used to analyse data with one follow-up measurement, but most researchers simply include the baseline measurement in a model as covariate, which is not sufficient and might greatly falsify the results. Our extended methods enable the correct inclusion of underlying true baseline values in several types of linear mixed models, both for univariate and multivariate outcome variables.

References

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle, *in* B. N. Petrov and F. Csaki (eds), *International Symposium on Information Theory*, Budapest: Akademia Kiado, pp. 267–281.

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test, *Journal of Multivariate Analysis* **88**(1): 190–206.

Bates, D. (2013). Linear mixed model implementation in `lme4`, `http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf`.

Bates, D. and DebRoy, S. (2004). Linear mixed models and penalized least squares, *Journal of Multivariate Analysis* **91**(1): 1–17.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability, *Monthly Weather Review* **78**(1): 1–3.

Bröcker, J. and Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper, *Weather and Forecasting* **22**(2): 382–388.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*, Cambridge University Press, Cambridge.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design, *The Annals of Statistics* **27**(1): 65–81.

Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd edn, Oxford University Press, Oxford.

Donohue, M., Overholser, R., Xu, R. and Vaida, F. (2011). Conditinal Akaike information under generalized linear and proportional hazards mixed models, *Biometrika* **98**(3): 685–700.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, Springer, New York.

Fahrmeir, L., Kneib, T. and Lang, S. (2007). *Regression - Modelle, Methoden und Anwendungen*, Springer, Heidelberg.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (eds) (2009). *Longitudinal Data Analysis*, Chapman & Hall, Boca Raton.

Frees, E. W. (2004). *Longitudinal and Panel Data*, Cambridge University Press, Cambridge.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models, *Statistics and Computing* **7**(1): 57–68.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association* **100**(470): 680–700.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation, *Journal of the American Statistical Association* **102**(477): 359–378.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society, Ser. B* **69**(2): 243–268.

Greven, S. and Kneib, T. (2010). On the behavior of marginal and conditional AIC in linear mixed models, *Biometrika* **97**(4): 773–789.

Harrison, L., Dunn, D., Green, H. and Copas, A. (2009). Modelling the association between patient characteristics and the change over time in a disease measure using observational cohort data, *Statistics in Medicine* **28**(26): 3260–3275.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*, Wiley, New York.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather and Forecasting* **15**(5): 559–570.

Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models, *Biometrika* **88**(2): 367–379.

Jiang, J., Rao, J., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection, *The Annals of Statistics* **36**(4): 1669–1692.

Kass, R. and Raftery, A. (1995). Bayes factors, *Journal of the American Statistical Association* **90**(430): 773–795.

Kneib, T. (2006). *Mixed Model Based Inference in Structured Additive Regression*, Verlag Dr. Hut, Munich.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *The Annals of Mathematical Statistics* **22**(1): 79–86.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**(4): 963–974.

Lian, H. (2012). A note on conditional Akaike information for Poisson regression with random effects, *Electronic Journal of Statistics* **6**: 1–9.

Liang, H., Wu, H. and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models, *Biometrika* **95**(3): 773–778.

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models, *Statistics in Medicine* **22**(10): 1649–1660.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer, New York.

Nguyen, T. and Jiang, J. (2012). Restricted fence method for covariate selection in longitudinal data analysis, *Biostatistics* **13**(2): 303–314.

Pauler, D. (1998). The Schwarz criterion and related methods for normal linear models, *Biometrika* **85**(1): 13–27.

Pauler, D., Wakefield, J. and Kass, R. (1999). Bayes factors and approximations for variance component models, *Journal of the American Statistical Association* **94**(448): 1242–1253.

Pawitan, Y. (2001). *In All Likelihood - Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford.

Pinheiro, J. and Bates, D. (2004). *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

Pu, W. and Niu, X. F. (2006). Selecting mixed-effects models based on a generalized information criterion, *Journal of Multivariate Analysis* **97**(3): 733–758.

Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory, *Monthly Weather Review* **130**(6): 1653–1660.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, Wiley, New York.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Ser. B* **39**(1): 44–47.

Székely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality, *Journal of Multivariate Analysis* **93**(1): 58–80.

Taylor, J. M. G. and Law, N. (1998). Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts?, *Statistics in Medicine* **17**(20): 2381–2394.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models, *Biometrika* **92**(2): 351–370.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer, New York.

Verbeke, G. and Molenberghs, G. (eds) (1997). *Linear Mixed Models in Practice*, Springer, New York.

Yu, D. and Yau, K. K. W. (2012). Conditional Akaike information criterion for generalized linear mixed models, *Computational Statistics and Data Analysis* **56**(3): 629–644.

PAPER I

Predictive crossvalidation for the choice of linear mixed-effects models with application to data from the Swiss HIV Cohort Study

Julia Braun, Leonhard Held & Bruno Ledergerber

Paper published in *Biometrics*, 2012, **68** (1), 53-61.

Biometrics 68, 53–61
March 2012

DOI: 10.1111/j.1541-0420.2011.01621.x

Predictive Cross-validation for the Choice of Linear Mixed-Effects Models with Application to Data from the Swiss HIV Cohort Study

Julia Braun,[1,*] Leonhard Held,[1] and Bruno Ledergerber[2]

[1]Biostatistics Unit, Institute for Social and Preventive Medicine, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland

[2]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, Rämistrasse 100, 8091 Zurich, Switzerland

[*]*email:* julia.braun@ifspm.uzh.ch

Summary. Model choice in linear mixed-effects models for longitudinal data is a challenging task. Apart from the selection of covariates, also the choice of the random effects and the residual correlation structure should be possible. Application of classical model choice criteria such as Akaike information criterion (AIC) or Bayesian information criterion is not obvious, and many versions do exist. In this article, a predictive cross-validation approach to model choice is proposed based on the logarithmic and the continuous ranked probability score. In contrast to full cross-validation, the model has to be fitted only once, which enables fast computations, even for large data sets. Relationships to the recently proposed conditional AIC are discussed. The methodology is applied to search for the best model to predict the course of CD4+ counts using data obtained from the Swiss HIV Cohort Study.

Key words: Cross-validation; Linear mixed-effects model; Predictive model choice; Proper scoring rules; Serial correlation.

1. Introduction

Model choice in normal linear mixed-effects models is not as straightforward as in normal linear models. Although in the latter case model choice reduces to the selection of relevant covariates, there are more challenges in a mixed-effects model: In addition to the choice of covariates included in the fixed effects, a decision on the type and number of random effects has to be made, as well as on the appropriate residual correlation structure if additional serial correlation is taken into account.

This greatly influences the concepts of model choice in general, and in particular the construction and applicability of the Akaike information criterion (AIC) and Bayesian information criterion (BIC). The AIC was suggested by Akaike (1973) and is defined as $\text{AIC} = -2\log L + 2p$, where p stands for the number of parameters in the model and L is the value of the likelihood at the parameter estimates. The BIC replaces the penalty $2p$ by $p\log(n)$, i.e., $\text{BIC} = -2\log L + p\log(n)$, where n represents the number of observations in the data set (see, e.g., Claeskens and Hjort, 2008).

These definitions have to be adapted for mixed-effects models. Pauler (1998) presents a generalization of the BIC, where the penalty term is adapted to the structure of unbalanced longitudinal data. However, this approach allows only a choice between fixed effects. A different attempt to perform selection of random effects is suggested by Pauler, Wakefield, and Kass (1999) and involves a boundary Laplace approximation, leading to an approximation to the BIC with an additional boundary correction term. Pu and Niu (2006) propose a generalized information criterion for model selection in linear mixed models, but the simulations in their article show that this criterion is not well suited for the choice of random effects.

The classical AIC can be applied in linear mixed-effects models, if inference concerning the population parameters is the main focus. Random effects are first integrated out to obtain the marginal likelihood and the corresponding marginal AIC (mAIC) criterion. However, Vaida and Blanchard (2005) argue that individual random effects are often of interest and introduce the conditional AIC (cAIC) as an alternative to mAIC. The effective degrees of freedom that determine the penalty term can be obtained as in Hodges and Sargent (2001). Liang, Wu, and Zou (2008) extend the work by Vaida and Blanchard (2005) and provide a general formula for the cAIC, where the variance components do not have to be known explicitly, because the uncertainty induced by estimating them can be taken into account. This ensures more reliable results; however, the penalty term can only be calculated numerically. Greven and Kneib (2010) show that the computational burden is immense, especially for large data sets, which causes problems when decisions between several models are desired.

As a general alternative to cAIC, we introduce a novel predictive model selection approach, which can be applied easily in linear mixed-effects models, even in the case of additional serial correlation. To our knowledge, no modification of cAIC has been developed in this case. It is generally recommended to choose the correlation structure based on the empirical variogram of suitable residuals (see, e.g., Diggle et al., 2002, p. 85) or by assessing the model fit (see Verbeke and

Molenberghs, 2000, p. 137). However, none of these methods provide a formal model choice criterion.

We focus on the cAIC due to its close relationship to the cross-validated logarithmic score (denoted by LS) that is presented later. Wang and Schaalje (2009) also state the importance of performing predictive model choice in linear mixed models, but they only use point predictions. In contrast, we incorporate the whole predictive distribution. Based on proper scoring rules (Gneiting and Raftery, 2007), we present a predictive cross-validation approach that can be applied to the selection of covariates, random effects, or the correlation structure. Geisser and Eddy (1979) proposed a similar predictive cross-validation approach to model selection, calling it "high structure selection." See also Krnjajic, Kottas, and Draper (2008) for a Bayesian approach relating the cross-validated LS to the deviance information criterion. To avoid refitting the model in each cross-validation step, we apply "mixed" cross-validation proposed by Marshall and Spiegelhalter (2003), which needs considerably less computation time than complete cross-validation. We also provide an empirical comparison with full cross-validation for different models fitted to data obtained from the Swiss HIV Cohort Study (SHCS).

This article is organized as follows. In Section 2, the linear mixed-effects model as well as the respective predictive distribution are presented. In Section 3, we review proper scoring rules and explain the cross-validation approach. An application to SHCS data is discussed in Section 4. Section 5 adds some general discussion.

2. Prediction in Linear Mixed-Effects Models

Linear mixed-effects models were presented by Laird and Ware (1982) and have been extended since then to incorporate various possible features of longitudinal data. They have the following general form: For each unit $i = 1, \dots, I$, there are J_i measurements at the timepoints t_{ij} for $j = 1, \dots, J_i$. The general model (see, e.g., Diggle et al., 2002) is written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{D}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

with fixed effects $\boldsymbol{\beta}$, random effects $\boldsymbol{b}_i$ and residuals $\boldsymbol{\epsilon}_i$, whereas $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are matrices of covariates for the fixed and random effects, respectively. The vector $\boldsymbol{D}_i$ takes into account serial correlation and is assumed to be an (independent) realization from a stationary Gaussian process with mean zero, variance ξ^2, and correlation function $\rho\,(|t_{ij} - t_{ik}|)$, thus incorporating the distance between the timepoints t_{ij} and t_{ik} of measurements j and k from the same individual i. The random effects $\boldsymbol{b}_i$ and the residuals $\boldsymbol{\epsilon}_i$ are mutually independent Gaussian random variables, with $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{J_i})$, and $\boldsymbol{b}_i \sim N(\boldsymbol{0},\, \boldsymbol{Q})$.

There are various choices for a suitable serial correlation function $\rho(|t_{ij} - t_{ik}|)$; among the most widely used are the exponential correlation model $\rho(|t_{ij} - t_{ik}|) = \exp\,(-\phi|t_{ij} - t_{ik}|)$ and the Gaussian correlation model $\rho(|t_{ij} - t_{ik}|) = \exp\,(-\phi|t_{ij} - t_{ik}|^2)$ for some value of $\phi > 0$. If a model with serial correlation is used, additional variation at a single timepoint, sometimes called nugget effect, is represented by the residual variance σ^2. However, estimation of σ^2 may be strongly model dependent if the data do not include duplicate measurements at the same time (Section 5.2.2, Diggle et al., 2002). If a model with serial correlation but without nugget effect shall be applied, the residual variance σ^2 is set to zero.

Maximum likelihood inference for linear mixed-effects models is based on the marginal distribution of $\boldsymbol{y}_i$. The mean of this distribution is simply

$$\mathrm{E}(\boldsymbol{y}_i) = \mathrm{E}(\mathrm{E}(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) = \mathrm{E}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta}. \tag{2}$$

The marginal covariance matrix is derived via

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{y}_i) &= \mathrm{E}(\mathrm{Cov}(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) + \mathrm{Cov}(\mathrm{E}(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) \\ &= \sigma^2\boldsymbol{I}_{J_i} + \xi^2\boldsymbol{\rho}(|\boldsymbol{u}_i|) + \boldsymbol{Z}_i\boldsymbol{Q}\boldsymbol{Z}_i^T,\end{aligned} \tag{3}$$

where $\boldsymbol{u}_i$ is a matrix that contains all time differences between measurements from the ith individual. For additional information on linear mixed-effects models for longitudinal data, see, e.g., Diggle et al. (2002) or Verbeke and Molenberghs (2000). Practical information on fitting those models can be found in Pinheiro and Bates (2004).

One reason for fitting statistical models is to make predictions for future observations. To evaluate the predictive properties of a specific model, it is necessary to calculate the parameters of the respective predictive distribution. To make a clear distinction between timepoints at which observations have already been made and future timepoints for which the predictive distribution is needed, the former are denoted by $\boldsymbol{t}_i = (t_{i1}, \dots, t_{ij}, \dots, t_{iJ_i})$ for individual $i = 1, \dots, I$ and measurement number $j = 1, \dots J_i$, whereas the "new" timepoints of the same individual are denoted by $\boldsymbol{s}_i = (s_{i1}, \dots, s_{ik}, \dots, s_{iK_i})$ for $k = 1, \dots, K_i$. Our interest lies in the predictive distribution of $\boldsymbol{y}_i(\boldsymbol{s}_i)$ conditional on the observations $\boldsymbol{y}_i(\boldsymbol{t}_i)$.

Derivation (see also Diggle et al., 2002, p. 111) of mean and covariance matrix of $\boldsymbol{y}_i(\boldsymbol{s}_i) \,|\, \boldsymbol{y}_i(\boldsymbol{t}_i)$ is based on the joint distribution of $\boldsymbol{Z}_i(\boldsymbol{s}_i)\boldsymbol{b}_i + \boldsymbol{D}_i(\boldsymbol{s}_i)$ and $\boldsymbol{y}_i(\boldsymbol{t}_i)$. Due to the mutual independence of $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$, this joint distribution is

$$\begin{aligned}&\begin{bmatrix}\boldsymbol{Z}_i(\boldsymbol{s}_i)\boldsymbol{b}_i + \boldsymbol{D}_i(\boldsymbol{s}_i) \\ \boldsymbol{y}_i(\boldsymbol{t}_i)\end{bmatrix} \\ &\quad\sim N\left\{\begin{bmatrix}\boldsymbol{0} \\ \boldsymbol{X}_i(\boldsymbol{t}_i)\boldsymbol{\beta}\end{bmatrix}, \begin{bmatrix}\boldsymbol{G}(\boldsymbol{s}_i, \boldsymbol{s}_i) & \boldsymbol{G}(\boldsymbol{s}_i, \boldsymbol{t}_i) \\ \boldsymbol{G}(\boldsymbol{t}_i, \boldsymbol{s}_i) & \sigma^2\boldsymbol{I}_{J_i} + \boldsymbol{G}(\boldsymbol{t}_i, \boldsymbol{t}_i)\end{bmatrix}\right\},\end{aligned} \tag{4}$$

where $\boldsymbol{G}\,(\boldsymbol{s}_i,\, \boldsymbol{t}_i)$ represents the covariance matrix $\mathrm{Cov}(\boldsymbol{Z}_i(\boldsymbol{s}_i)\,\boldsymbol{b}_i + \boldsymbol{D}_i(\boldsymbol{s}_i),\, \boldsymbol{Z}_i(\boldsymbol{t}_i)\boldsymbol{b}_i + \boldsymbol{D}_i(\boldsymbol{t}_i))$, which contains the elements of $\boldsymbol{Q}$, $\xi^2\ \boldsymbol{\rho}\ (|\boldsymbol{u}_i|)$, and σ^2.

Having determined the joint distribution of $\boldsymbol{Z}_i(\boldsymbol{s}_i)\boldsymbol{b}_i + \boldsymbol{D}_i(\boldsymbol{s}_i)$ and $\boldsymbol{y}_i(\boldsymbol{t}_i)$, standard properties of the multivariate normal distribution can be used to derive the conditional distribution of $\boldsymbol{Z}_i(\boldsymbol{s}_i)\boldsymbol{b}_i + \boldsymbol{D}_i(\boldsymbol{s}_i) \,|\, \boldsymbol{y}_i(\boldsymbol{t}_i)$, in particular

$$\begin{aligned}&\mathrm{E}(\boldsymbol{Z}_i(\boldsymbol{s}_i)\boldsymbol{b}_i + \boldsymbol{D}_i(\boldsymbol{s}_i) \mid \boldsymbol{y}_i(\boldsymbol{t}_i)) \\ &\quad= \boldsymbol{G}(\boldsymbol{s}_i, \boldsymbol{t}_i)\{\sigma^2\boldsymbol{I}_{J_i} + \boldsymbol{G}(\boldsymbol{t}_i, \boldsymbol{t}_i)\}^{-1}(\boldsymbol{y}_i(\boldsymbol{t}_i) - \boldsymbol{X}_i(\boldsymbol{t}_i)\boldsymbol{\beta}),\end{aligned} \tag{5}$$

and

$$\begin{aligned}&\mathrm{Cov}(\boldsymbol{Z}_i(\boldsymbol{s}_i)\boldsymbol{b}_i + \boldsymbol{D}_i(\boldsymbol{s}_i) \mid \boldsymbol{y}_i(\boldsymbol{t}_i)) \\ &\quad= \boldsymbol{G}(\boldsymbol{s}_i, \boldsymbol{s}_i) - \boldsymbol{G}(\boldsymbol{s}_i, \boldsymbol{t}_i)\{\sigma^2\boldsymbol{I}_{J_i} + \boldsymbol{G}(\boldsymbol{t}_i, \boldsymbol{t}_i)\}^{-1}\boldsymbol{G}(\boldsymbol{t}_i, \boldsymbol{s}_i).\end{aligned} \tag{6}$$

Note that $\boldsymbol{X}_i(\boldsymbol{s}_i)\boldsymbol{\beta}$ and, respectively, $\sigma^2\boldsymbol{I}_{K_i}$ have to be added to (5) and (6) to obtain the corresponding moments of the predictive distribution of $\boldsymbol{y}_i(\boldsymbol{s}_i) \,|\, \boldsymbol{y}_i(\boldsymbol{t}_i)$.

In practice, the regression coefficients and variance parameters are replaced by their estimates $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{Q}}, \hat{\xi}^2, \hat{\rho}$, and $\hat{\sigma}^2$, which also determine the elements of $\boldsymbol{G}(\boldsymbol{s}_i, \boldsymbol{t}_i)$. The uncertainty of these estimates is then ignored, which, according to Diggle et al. (2002, p. 112), should not be problematic if the sample size is reasonably large. Note that (5) and (6) can be used for multivariate predictions, but from now on we will focus on univariate predictions, i.e., $K_i = 1$.

3. Predictive Cross-validation

Several suggestions for evaluating the predictive accuracy of distinct models can be found in the literature. Most of these methods compare the point prediction with the true value using the mean squared error (e.g., Lee, 1988). Taylor and Law (1998) additionally use coverage, average bias, and percentage of underestimation to validate predictions. Alternatively the mean relative squared deviation, the mean absolute deviation, or the mean absolute relative deviation could be used (Keramidas and Lee, 1990).

These techniques do not take into account the properties of the whole predictive distribution and are therefore not sufficient. Only the distance between the point prediction and the observation is considered, but the variance of the predictive distribution is omitted. Using formulae (5) and (6), measures that take into account the whole predictive distribution can be obtained, thus incorporating much more information of the statistical model used to generate predictions.

3.1 *Proper Scoring Rules*

We use proper scoring rules to compare a predictive distribution Y with the observed value y_{obs}. They allow calculating a numerical value for each of the competing models, which can then be compared to determine the best predictive model. They are typically positively oriented, so that a larger value denotes the better model. Proper scoring rules are defined as follows: The expected value of a proper score becomes maximal if the observed value is a realization from the assumed predictive distribution. A score is strictly proper if this maximum is unique. Propriety is an essential property of a scoring rule, ensuring that it addresses calibration and sharpness simultaneously (Winkler, 1996). Note that the mean of proper scores remains proper (Gneiting, Balabdaoui, and Raftery, 2007), which allows us to summarize the scores of the predictions for several measurements.

A very common score is the LS, which evaluates the log predictive density $f_Y(y)$ at the observed value y_{obs}:

$$\mathrm{LS}(Y, y_{obs}) = \log f_Y(y_{obs}). \tag{7}$$

In the case of a univariate normal predictive distribution with mean μ and variance σ^2, the LS takes the form

$$\mathrm{LS}(Y, y_{obs}) = -\frac{1}{2}(\log(2\pi) + \log(\sigma^2) + (y_{obs} - \mu)^2/\sigma^2).$$

As the predictive distribution is only evaluated at the actual observation y_{obs}, the LS is a local score, which ignores all other values of $f_Y(y)$. For this reason the LS is regularly criticized as being too limited.

Another proper scoring rule for univariate predictive distributions is the continuous ranked probability score (CRPS). It is the integral over all possible thresholds r of the Brier score (Brier, 1950), a well-known score for binary predictions, and takes the form

$$\mathrm{CRPS}(Y, y_{obs}) = -\int_{-\infty}^{\infty} (P_Y(Y \le r) - 1(y_{obs} \le r))^2 \, dr. \tag{8}$$

In contrast to the LS, the CRPS rewards high probabilities for values close to the observed one, whereas lower probabilities of these values are penalized. Consequently, this score is sensitive to distance and not a local score, as it takes into account how close a predicted value is to the observed value (see also Hersbach, 2000; Gneiting and Raftery, 2007).

The integral in (8) is often difficult to obtain; however, Gneiting and Raftery (2007) showed that the CRPS can be written as

$$\mathrm{CRPS}(Y, y_{obs}) = \frac{1}{2} E_{f_Y} |Y - Y'| - E_{f_Y} |Y - y_{obs}|, \tag{9}$$

where Y and Y' are independent realizations from the predictive density f_Y. If the predictive distribution is univariate Gaussian, an explicit form of the CRPS can be obtained from (9):

$$\mathrm{CRPS}(Y, y_{obs}) = \sigma \left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{y_{obs} - \mu}{\sigma}\right) - \frac{y_{obs} - \mu}{\sigma}\left(2\Phi\left(\frac{y_{obs} - \mu}{\sigma}\right) - 1\right)\right],$$

where φ and Φ denote the probability density function and the cumulative distribution function of a standard Gaussian variable.

There is no automatic choice of a proper scoring rule to be used in any given situation, unless there is a unique and clearly defined underlying decision problem. However, in many types of situations it may be appropriate to use a variety of diagnostic tools and scores, to take advantage of their differing emphases and strengths.

3.2 *Predictive Cross-validation*

The idea of the cross-validation approach presented here is based on the concept of mixed predictive model checks by Marshall and Spiegelhalter (2003), an alternative to classical cross-validation. They suggest fitting the model only once to the full data, but the individual random effects as well as the true measurement are ignored and newly generated in the forecasting process of each cross-validation step. In this case the danger of conservatism is reduced considerably, because the omitted observation does not influence the random effects directly, but only via the hyperparameters. This approach has been also used by Riebler and Held (2010) and Held, Schrödle, and Rue (2010).

We suggest a similar cross-validation approach: First, the chosen model is fit to the whole data set. One of the observations is then left out, corresponding to a scalar s_i in (5), so that $\boldsymbol{t}_i$ is reduced by one component. Calculation of the predictive distribution only involves the hyperparameters of the random effects and the fixed effects parameters, so it is easy to calculate the LS and the CRPS for this specific observation. This is repeated once for each observation in the data set, so that in the end, the cross-validated mean scores $\overline{\mathrm{LS}}_{CV}$ and $\overline{\mathrm{CRPS}}_{CV}$ are obtained.

The difference between full cross-validation and our approach becomes clearer if we have a look at the respective

versions of formulae (5) and (6). If full cross-validation is conducted, the model is fit once for each measurement left out. In our notation, the subscript $-j$ denotes values estimated from a model without measurement j of an arbitrary individual i. This leads to

$$\begin{aligned}\mathrm{E}(y_i(s_{ij}) \,|\, \boldsymbol{y}_i(\boldsymbol{t}_{i,-j})) = {} & \boldsymbol{X}_i(s_{ij})\hat{\boldsymbol{\beta}}_{-j} + \hat{\boldsymbol{G}}_{-j}(s_{ij}, \boldsymbol{t}_{i,-j}) \\ & \times \{\hat{\sigma}^2_{-j}\boldsymbol{I}_{J_i-1} + \hat{\boldsymbol{G}}_{-j}(\boldsymbol{t}_{i,-j}, \boldsymbol{t}_{i,-j})\}^{-1} \\ & \times (\boldsymbol{y}_i(\boldsymbol{t}_{i,-j}) - \boldsymbol{X}_i(\boldsymbol{t}_{i,-j})\hat{\boldsymbol{\beta}}_{-j}),\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(y_i(s_{ij}) \,|\, \boldsymbol{y}_i(\boldsymbol{t}_{i,-j})) = {} & \hat{\sigma}^2_{-j} + \hat{\boldsymbol{G}}_{-j}(s_{ij}, s_{ij}) - \hat{\boldsymbol{G}}_{-j}(s_{ij}, \boldsymbol{t}_{i,-j}) \\ & \times \{\hat{\sigma}^2_{-j}\boldsymbol{I}_{J_i-1} + \hat{\boldsymbol{G}}_{-j}(\boldsymbol{t}_{i,-j}, \boldsymbol{t}_{i,-j})\}^{-1} \\ & \times \hat{\boldsymbol{G}}_{-j}(\boldsymbol{t}_{i,-j}, s_{ij}).\end{aligned}$$

In contrast, our cross-validation approach involves only one model fit at the beginning, so that the estimated hyperparameters $\hat{\boldsymbol{G}}_{-j}$ and $\hat{\sigma}^2_{-j}$ as well as the fixed effects $\hat{\boldsymbol{\beta}}_{-j}$ are replaced by their estimates based on all data.

3.3 *Justification and Implementation of the Predictive Cross-validation Approach*

In a landmark paper, Stone (1977) showed that in a conventional parametric statistical model (without random effects) for n independent observations, the AIC is asymptotically equivalent to the cross-validated mean LS: AIC $\doteq -2n\overline{\mathrm{LS}}_{CV}$. See Section 13.6 of Pawitan (2001) for a useful summary. In the linear model (with known variance σ^2), the penalty $2p$, where p is the number of regression coefficients, can be written as $2\,\mathrm{tr}(\boldsymbol{H})$, where $\boldsymbol{H}$ is the classical hat matrix, projecting observations $\boldsymbol{y}$ to $\hat{\boldsymbol{y}}$. The trace of the hat matrix, often called the (effective) degrees of freedom, is also used as penalty term in nonparametric regression models, for example in generalized additive (Hastie and Tibshirani, 1990) or kernel regression (Herrmann, 2000) models.

We now sketch that $\overline{\mathrm{LS}}_{CV}$ in linear mixed-effects models is asymptotically equivalent to cAIC, defined as minus twice the conditional log likelihood at the parameter estimates plus twice the effective degrees of freedom (Vaida and Blanchard, 2005). To derive the cAIC in linear mixed-effects models without serial correlation, Vaida and Blanchard (2005) follow in Appendix 1 Hodges and Sargent (2001) and rewrite the linear mixed-effects model as a standard linear model without random effects, but with added "pseudodata." The random effects are now treated as fixed, which exactly corresponds to the first likelihood contribution in the cAIC approach introduced by Vaida and Blanchard (2005). Now Hodges and Sargent (2001) define the trace of the corresponding hat matrix (ignoring the pseudodata) as the effective number of degrees, and this is exactly the penalty derived for the cAIC, see Vaida and Blanchard (2005, Appendix 2). So the cAIC can be viewed as an ordinary AIC in this artificial linear model with pseudodata. The cross-validated mean LS, known to be asymptotically equivalent to AIC, in the above linear model with pseudodata now equals our proposed leave-one-observation-out cross-validation approach, because the leave-one-out prediction of the pseudodata have zero variance and hence make no contribution to $\overline{\mathrm{LS}}_{CV}$. This suggests the asymptotic equality of $\overline{\mathrm{LS}}_{CV}$ and $-\mathrm{cAIC}/(2\sum_i J_i)$, which is further confirmed empirically in Section 4.

We note that for models with additional serial collection a similar cross-validation procedure is suggested in van der Linde (1994) in the context of spline smoothing with dependent errors. Note that as a general alternative to the LS, we also examined the usability of the CRPS for the same task.

For comparison, we have also derived the effective degrees of freedom for models with additional serial correlation based on the Vaida and Blanchard (2005) approach. Details can be found in Web Appendices A and B. The calculation of the corrected cAIC is, however, difficult: An analytic representation for models with serial correlation has not yet been deducted and would require that all variance components have to be known (Kneib, 2010, personal communication). The alternative numerical calculation is practically impossible for large data sets, because it involves fitting at least $\sum_i J_i$ models and would therefore take much too long. This has also been illustrated in Greven and Kneib (2010).

For the analyses in this article, the `lme` function from the `R` package `nlme` was used for fitting linear mixed-effects models with serial correlation. Similar procedures are implemented in most standard statistical software, e.g., the procedure `PROC MIXED` in `SAS` (see Verbeke and Molenberghs, 1997, 2000). Although there already exists a `predict.lme` function in `R`, this function is not sufficient for our purposes, because it provides only the predictions itself (i.e., the predictive mean), but not the respective predictive variance, and does not take into account serial correlation. Therefore, we implemented functions that provide both the predictive mean and the predictive covariance for linear mixed-effects models with random effects and serial correlation with or without a nugget effect as well as a function to conduct our proposed cross-validation approach. This function and all its subfunctions can be found in the Supplementary Web Material, along with short explanations.

4. Application: HIV Progression in CD4+ Counts

We use a sample from the SHCS (`www.shcs.ch`) to illustrate our proposed method. The SHCS is a national research project of HIV positive adults. We investigate the course of the patients' CD4+ cell counts after starting antiretroviral therapy. CD4+ lymphocytes are a marker of the immune system and the main target of the HIV. In most individuals the HIV infection is not detected until CD4+ cells have declined considerably. As a consequence, the antiretroviral therapy is usually started at relatively low CD4+ cell counts. As soon as the HIV-1 viral load is suppressed by adequate antiretroviral treatment, CD4+ cells begin to rise and often reach levels close to normal. As a sudden decrease of the CD4+ counts usually is a hint for therapy failure requiring a regimen change, the importance of predicting the further course of the CD4+ counts is obvious.

It is a well-known fact that predictions in the context of HIV are particularly challenging, because the viral load and its effect on the immune system can change so fast. Furthermore, one may wish to choose the best model, and for this task, we applied our proposed cross-validation procedure. Selection of patients was performed as in Wolbers et al. (2007): Treatment-naive patients starting highly active antiretroviral therapy (HAART) with a baseline measurement and at

least two subsequent measurements of CD4+ cells were included. We excluded patients in whom viral load remained detectable for more than 9 months of therapy as this treatment failure is likely due to adherence problems. To permit the inclusion of baseline treatment information in our models, the data were censored as soon as the treatment had to be changed for the first time. Our data set consists of 6436 CD4+ measurements of 885 patients. The number of measurements per patient ranges from 2 to 38, and we model the square root of the CD4+ counts. See Web Figure 1, which shows the course of the CD4+ counts for some patients from our data set.

Potentially influential variables for the course of the CD4+ counts are time since HAART start, sex, age at HAART start, AIDS-defining illnesses prior to HAART start, HIV transmission category (via blood, heterosexual contacts, injecting drug use, homosexual contacts, perinatal, or other modes of transmission), time between HIV infection and HAART start, coinfections with hepatitis B or C, square root of the baseline CD4+ measurement (i.e., the last measurement before HAART start), $\log_{10}$ of the baseline viral load measurement, type of therapy regimen (nonnucleoside reverse-transcriptase inhibitors, ritonavir-boosted protease inhibitor, single protease inhibitor, or other), and the nucleoside pair used in the medication (any with tenofovir, any with stavudine and no tenofovir, lamivudine/zidovudine, or other).

We considered models with just a random intercept (denoted by "ri") as well as models with additional random slope (denoted by "ris"). Moreover, we looked at models without serial correlation, with exponential ("Exp") and with Gaussian ("Gauss") correlation structure. When fitting the models with serial correlation structure, we applied both models with and without a nugget effect (denoted by "nug" or "nonug," respectively). Combining these three aspects results in ten distinct model structures. Within each of these structures, models with five different combinations of covariates are compared. We consider two very sparse models with just time or, additionally, time squared as influential variables (denoted by M1 and M2), although a model with all possible variables (M5) is also included. Based on initial studies of model fit, we additionally looked at a model with time, time squared, and the two baseline variables (M3), whereas model M4 contains the variables of model M3 as well as AIDS at baseline, time between infection and start of the therapy, type of therapy regimen, and the nucleoside pair. The analyses in this article are done using maximum likelihood estimation, so that the (conditional) AIC values can also be computed for comparison.

The cross-validation procedure described in Section 3 was applied to all these models. The resulting mean scores are summarized in Table 1, where the best score of each column is printed in bold face and the best score of all is marked using italics. Concerning the choice of the general model structure within the same set of covariates, we see that both scores prefer a model with exponential correlation to the models without any or with Gaussian correlation structure. The model with random intercept, random slope, and nugget effect is preferred to the one with just random intercept or no nugget effect in all cases.

Focusing on the choice of covariates within the preferred model structure shows that both the LS and the CRPS appear to prefer model M3. However, the difference from models M4 and M5 is extremely small in both cases, whereas the difference from models M1 and M2 is noticeable. It can thus be concluded that the main information that contributes to better forecasts is mainly contained in the baseline CD4+ and viral load measurements whereas there is only a slight improvement of the predictive distribution when additional variables are added to the model.

Table 1
Mean LS, CRPS, and mean squared error of the Marshall–Spiegelhalter cross-validation procedure

	M1	M2	M3	M4	M5
Mean LS:					
ri	−2.392	−2.347	−2.339	−2.339	−2.338
ris	−2.294	−2.279	−2.269	−2.269	−2.268
Exp, ri, nug	−2.262	−2.255	−2.248	−2.247	−2.246
Exp, ris, nug	**−2.261**	**−2.254**	***−2.243***	**−2.244**	**−2.244**
Gauss, ri, nug	−2.272	−2.263	−2.255	−2.255	−2.254
Gauss, ris, nug	−2.287	−2.272	−2.262	−2.262	−2.262
Exp, ri, nonug	−2.326	−2.300	−2.293	−2.293	−2.292
Exp, ris, nonug	−2.277	−2.265	−2.255	−2.254	−2.254
Gauss, ri, nonug	−2.383	−2.339	−2.331	−2.331	−2.330
Gauss, ris, nonug	−2.287	−2.272	−2.262	−2.262	−2.262
Mean CRPS:					
ri	−1.426	−1.363	−1.353	−1.353	−1.352
ris	−1.297	−1.275	−1.262	−1.262	−1.262
Exp, ri, nug	−1.251	−1.240	−1.231	−1.230	−1.229
Exp, ris, nug	**−1.249**	**−1.239**	***−1.226***	**−1.227**	**−1.227**
Gauss, ri, nug	−1.264	−1.250	−1.241	−1.240	−1.239
Gauss, ris, nug	−1.290	−1.268	−1.255	−1.255	−1.255
Exp, ri, nonug	−1.303	−1.278	−1.271	−1.270	−1.270
Exp, ris, nonug	−1.267	−1.252	−1.240	−1.240	−1.240
Gauss, ri, nonug	−1.413	−1.352	−1.343	−1.343	−1.342
Gauss, ris, nonug	−1.290	−1.268	−1.255	−1.255	−1.255
Mean squared error score:					
ri	−6.905	−6.331	−6.243	−6.239	−6.232
ris	−5.757	−5.589	−5.480	−5.477	−5.471
Exp, ri, nug	−5.418	−5.329	−5.247	−5.240	−5.233
Exp, ris, nug	**−5.401**	**−5.322**	***−5.206***	**−5.219**	**−5.214**
Gauss, ri, nug	−5.530	−5.416	−5.332	−5.325	−5.317
Gauss, ris, nug	−5.669	−5.505	−5.399	−5.396	−5.390
Exp, ri, nonug	−5.801	−5.563	−5.504	−5.499	−5.490
Exp, ris, nonug	−5.470	−5.351	−5.250	−5.248	−5.243
Gauss, ri, nonug	−6.738	−6.200	−6.118	−6.115	−6.107
Gauss, ris, nonug	−5.669	−5.505	−5.399	−5.396	−5.390

If we compare models with and without a nugget effect, but having the same random effects and serial correlation structure, we see that the models with nugget effect are generally preferred. Although in the cases with exponential correlation the difference is not too strong, the models with Gaussian correlation, random intercept, and nugget effect are considerably worse than the equivalent models without nugget effect. On the other hand, the same models with additional random slope show practically the same scores in both cases. This can be explained by the fact that for these models, the nugget parameter is estimated to be almost zero and therefore leads to no difference between models with and without nugget effect.

If we have a look at the lower part of Table 1, the mean LS and CRPS can be compared with the cross-validated mean

Table 2
Conditional and corrected cAIC (transformed)

	M1	M2	M3	M4	M5
Conventional cAIC:					
ri	−2.393	−2.347	−2.340	−2.340	−2.339
ris	−2.285	−2.271	−2.262	−2.262	−2.262
Exp, ri, nug	−2.447	−2.429	−2.365	−2.364	−2.362
Exp, ris, nug	−2.388	−2.373	−2.336	−2.346	−2.347
Gauss, ri, nug	−2.427	−2.386	−2.360	−2.358	−2.357
Gauss, ris, nug	−2.282	−2.268	−2.260	−2.260	***−2.259***
Exp, ri, nonug	−2.387	−2.345	−2.335	−2.334	−2.334
Exp, ris, nonug	−2.292	−2.276	−2.266	−2.266	−2.266
Gauss, ri, nonug	−2.389	−2.344	−2.336	−2.336	−2.336
Gauss, ris, nonug	**−2.282**	**−2.268**	**−2.260**	**−2.260**	−2.260
Corrected cAIC:					
ri	−2.393	−2.347	−2.340	−2.340	−2.339
ris	**−2.267**	**−2.254**	***−2.247***	**−2.248**	**−2.249**

squared error. We see that the mean squared error, which does not take into account the predictive variance but is also a proper (however, not strictly proper) score, chooses the same model as the two strictly proper scores. This indicates that in this case, taking into account the predictive variance does not alter the decision on the best model.

The cAIC (Vaida and Blanchard, 2005) for the models without serial correlation can be found in Table 2. Note that the values have been divided by $-2\sum J_i$ to make them comparable to the mean LS values in Table 1. They show that these models are ranked in a similar order as is the case with the mean LS or CRPS. However, for models with serial correlation, where the appropriate correction of the cAIC has to be applied, the ranking of the models is quite different, and a different model is chosen as best suited for prediction. This time, models with Gaussian correlation and random intercept and slope are preferred to models with exponential correlation structure. Apart from that, models with no nugget effect are chosen in many cases, whereas this is not the case when proper scoring rules are used.

In the lower part of Table 2, the values of the corrected cAIC according to Greven and Kneib (2010) are shown for the models without serial correlation. For models with random intercept, they are extremely close to the cAIC from above, whereas there is quite a difference when models with additional random slope are concerned. However, the ordering stays more or less the same.

This is also illustrated in Figure 1, where the values of the LS are plotted against the values of the cAIC, which were transformed analogously to $-\text{cAIC}/(2\sum J_i)$. The models with random intercept only are shown on the left, the models with additional random slope on the right side. Note that in the case with random slope and Gaussian correlation, the estimated nugget effect is essentially zero so the results are identical to those obtained from the corresponding model without nugget effect. The agreement is quite good if no serial correlation is involved, whereas it becomes worse for models with serial correlation, but without nugget effect. In the case of an additional nugget effect, the LS and cAIC values differ strongly. Apart from that, the agreement both with and without a nugget effect is better for models with Gaussian correlation than with exponential correlation structure.

We also computed empirical variograms of the residuals for selected models (all with covariate combination M3), which are shown in Figure 2. The variograms in the left column belong to model M3 with exponential correlation structure and random intercept and slope, above with nugget effect, below without nugget effect. This exploratory analysis suggests choosing the model with nugget effect, which is consistent with the choice based on proper scoring rules. The cAIC, however, prefers models without nugget effect.

On the right side of Figure 2, we see the variograms of model M3 with Gaussian correlation and random intercept and slope. As the nugget effect is estimated to be almost zero, both the model with and without nugget effect have essentially the same empirical semivariogram. This variogram is quite similar to the one on the left, so that one would probably select the slightly simpler model with exponential instead of Gaussian correlation structure if the choice was based on empirical assessment only.

To compare our cross-validation with a real cross-validation approach, we also conducted a full cross-validation for the models. Figure 3 shows the comparison of the results for the LS and the CRPS. The mean scores of the full cross-validation were very similar to our results, they changed only at the third or fourth decimal place. Our cross-validation approach tended to show slightly higher results than the full leave-one-out approach, the expected behavior. Only in two cases the mean LS of the full cross-validation was lower than with our procedure. Both were models with Gaussian correlation and random intercept and slope, i.e., models that tended to be numerically unstable as also suggested by the variograms in Figure 2. In all other cases the differences were very small, so they can be safely ignored. In particular, the order of the results and thus the choice of the best model did not change.

Concerning the running time of the calculations, our cross-validation approach took less than 10 minutes per model, whereas a full cross-validation approach needs almost 2 hours even for the simplest model without serial correlation, and is prolonged considerably to a duration of up to 4 days if a serial correlation structure is added. In summary, the cross-validation approach presented here seems to work sufficiently well for the purpose of comparing linear mixed-effects models with respect to their predictive abilities. Model choice based on the cross-validated proper scoring rules is a viable alternative to the cAIC, in particular if serial correlation is included.

5. Summary and Discussion

In this article, we proposed a novel predictive approach to model selection of linear mixed-effects models with serial correlation. The cross-validated LS and CRPS can be calculated easily and provide a useful alternative to the cAIC. We note that the individual LS and CRPS values can also be used to assess calibration of a single model (Held, Rufibach, and Balabdaoui, 2010). We have applied these techniques in the context of this article but have omitted a description of the results due to space limitations.

Our proposed predictive cross-validation approach has the advantage that the model has to be fitted just once and not in every cross-validation step. This saves a considerable amount

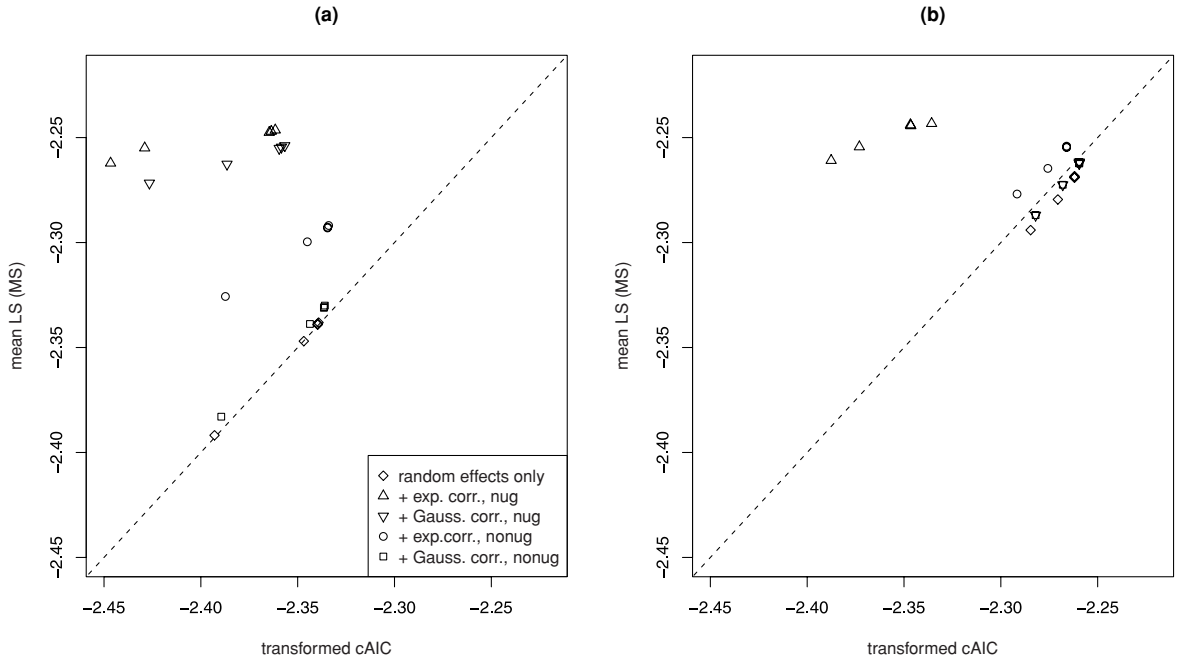

Figure 1. Mean cross-validated LS (Marshall–Spiegelhalter) versus transformed cAIC; (a) models with random intercept only; (b) models with additional random slope.

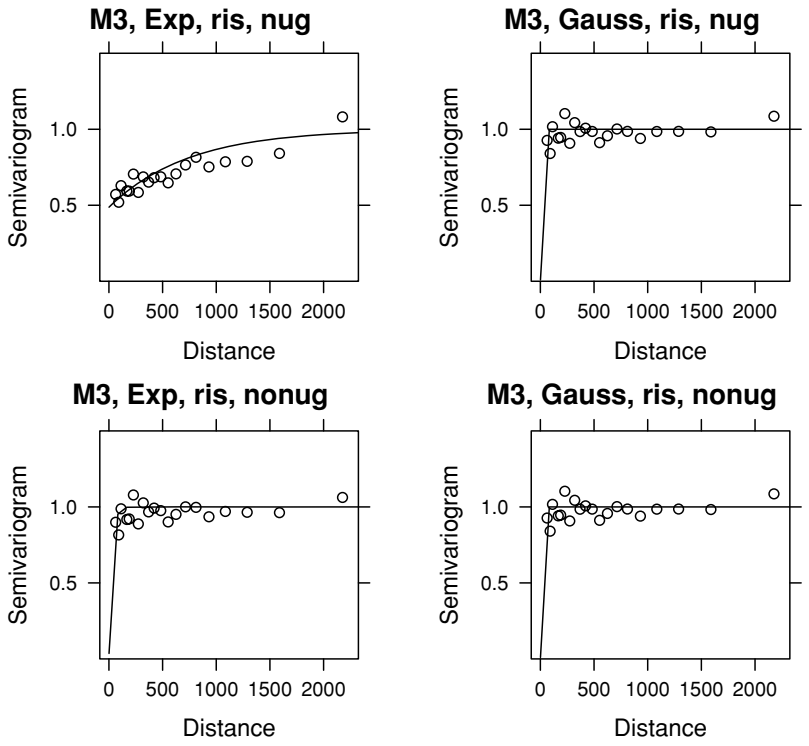

Figure 2. Variograms of selected models.

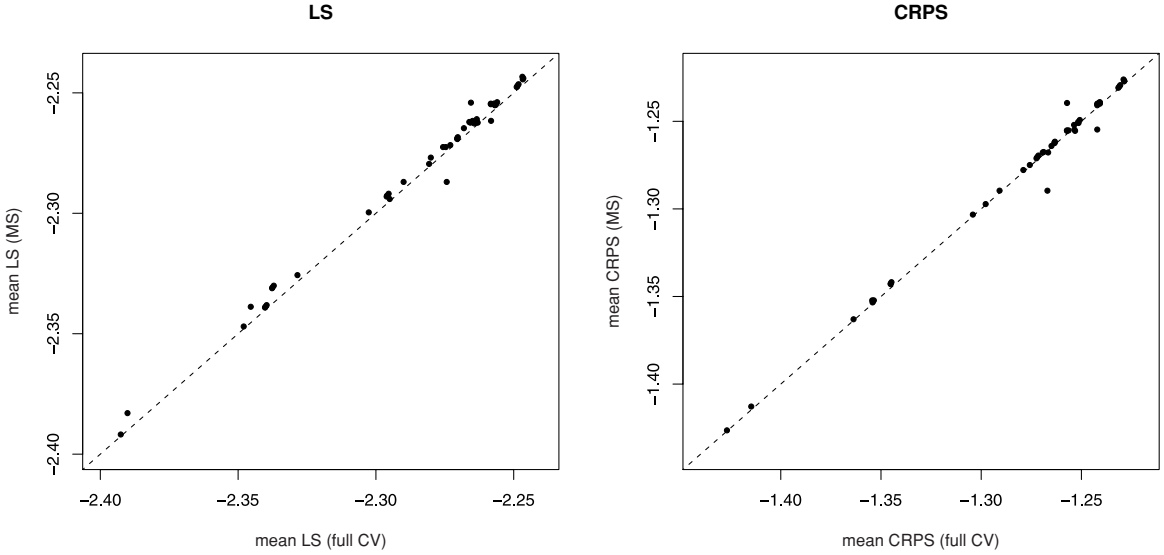

Figure 3. Comparison of our approach with a full cross-validation.

of time, especially in more complicated linear mixed-effects models with random effects and serial correlation. Empirical evidence suggests that this cross-validation approach is a good approximation of a true leave-one-out approach, but this of course depends on the nature of the concrete problem and data set. Moreover, this approach provides an asymptotically equivalent alternative to the cAIC, which has not been developed so far for models with serial correlation. Analogously, Dawid (1984) shows that the sum of one-step-ahead LSs is asymptotically equivalent to the BIC. This could be a promising alternative for model choice in linear mixed-effects models, which we will consider in our future work.

It is a well-known fact that predictions in the context of HIV are particularly challenging, because the viral load and its effect on the immune system can change so fast. Furthermore, one may wish to choose the best model, and for this task, our cross-validation procedure has shown to be useful.

6. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 3 and 4 as well as the `R` functions implementing our proposed method are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

We thank the SHCS for kindly providing the data set, Daniel Sabanés Bové for many helpful discussions, as well as an associate editor and two referees for helpful comments on earlier versions of this article.

References

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. In *International Symposium on Information Theory*, B. N. Petrov and F. Csaki (eds), pp. 267–281. Budapest, Hungary: Akademia Kiado.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78,** 1–3.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge, U.K.: Cambridge University Press.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A* **147,** 278–292.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd edition. Oxford: Oxford University Press.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74,** 153–160.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102,** 359–378.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B* **69,** 243–268.

Greven, S. and Kneib, T. (2010). On the behavior of marginal and conditional AIC in linear mixed models. *Biometrika* **97,** 773–789.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Held, L., Rufibach, K., and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics* **66,** 1295–1305.

Held, L., Schrödle, B., and Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In *Statistical Modelling and Regression Structures—Festschrift in Honour of Ludwig Fahrmeir*, T. Kneib and G. Tutz (eds), pp. 91–110. Heidelberg, Germany: Physica-Verlag.

Herrmann, E. (2000). Variance estimation and bandwidth selection for kernel regression. In *Smoothing and Regression: Approaches, Computation, and Application*, M. G. Schimek (ed.), pp. 71–107. New-York: John Wiley & Sons.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting* **15,** 559–570.

Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* **88,** 367–379.

Keramidas, E. M. and Lee, J. C. (1990). Forecasting technological substitutions with concurrent short time series. *Journal of the American Statistical Association* **85,** 625–632.

Krnjajic, M., Kottas, A., and Draper, D. (2008). Parametric and nonparametric Bayesian model specification: A case study involving models for count data. *Journal of Computational Statistics and Data Analysis* **52,** 2110–2128.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

Lee, J. C. (1988). Prediction and estimation of growth curves with special covariance structures. *Journal of the American Statistical Association* **83,** 432–440.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* **95,** 773–778.

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine* **22,** 1649–1660.

Pauler, D. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85,** 13–27.

Pauler, D., Wakefield, J., and Kass, R. (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association* **94,** 1242–1253.

Pawitan, Y. (2001). *In All Likelihood—Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Pinheiro, J. and Bates, D. (2004). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Pu, W. and Niu, X. F. (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of Multivariate Analysis* **97,** 733–758.

Riebler, A. and Held, L. (2010). The analysis of heterogeneous time trends in multivariate age-period-cohort models. *Biostatistics* **11,** 57–69.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* **39,** 44–47.

Taylor, J. M. G. and Law, N. (1998). Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? *Statistics in Medicine* **17,** 2381–2394.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92,** 351–370.

van der Linde, A. (1994). On cross-validation for smoothing splines in the case of dependent observations. *Australian and New Zealand Journal of Statistics* **36,** 67–73.

Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models in Practice*. New York: Springer.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wang, J. and Schaalje, G. B. (2009). Model selection for linear mixed models using predictive criteria. *Communications in Statistics—Simulation and Computation* **38,** 788–801.

Winkler, R. (1996). Scoring rules and the evaluation of probabilities. *Test* **5,** 1–60.

Wolbers, M., Battegay, M., Hirschel, B., Furrer, H., Cavassini, M. B. H., Vernazza, P. L., Bernasconi, E., Kaufmann, G., Bucher, H. C., and the Swiss HIV Cohort Study. (2007). CD4+ T-cell count increase in HIV-1-infected patients with suppressed viral load within 1 year after start of antiretroviral therapy. *Antiviral Therapy* **12,** 889–897.

Received February 2010. *Revised March* 2011.
Accepted April 2011.

Web-based Supplementary Materials for "Predictive cross-validation for choice of linear mixed-effects models with application to data from the Swiss HIV Cohort Study" by Julia Braun, Leonhard Held and Bruno Ledergerber

March 28, 2011

Web Appendix A

Bias correction term for AIC in linear mixed-effects models without serial correlation

The bias correction term BC for a linear mixed-effects model where the residual variance σ^2 and the covariance matrix of the random effects $\boldsymbol{W}$ are known is deducted in Appendix 2 of Vaida and Blanchard (2005). This deduction works as follows:

Let $g(y|b)$ be the conditional likelihood of the model, and $\boldsymbol{y}^*$ be the prediction data set such that $\boldsymbol{y}^*$ and $\boldsymbol{y}$ are independent conditional on the true random effects. The number of measurements per individual n_i goes to infinity, so that we skip the index i. Apart from that, $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ and $\boldsymbol{y}^* = \boldsymbol{\mu} + \boldsymbol{\epsilon}^*$, as well as $\boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{Zb}$. Finally, $\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^* \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$.

$$
\begin{aligned}
BC &= \mathrm{E}_y \log g\{\boldsymbol{y}|\hat{\boldsymbol{b}}(\boldsymbol{y})\} - \mathrm{E}_y \mathrm{E}_{y^*} \log g\{\boldsymbol{y}^*|\hat{\boldsymbol{b}}(\boldsymbol{y})\} = \\
&= \mathrm{E}_y\{-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2\} - \mathrm{E}_y \mathrm{E}_{y^*}\{-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||\boldsymbol{y}^* - \hat{\boldsymbol{y}}||^2\} = \\
&= \mathrm{E}(||\boldsymbol{y}^* - \hat{\boldsymbol{y}}||^2 - ||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2)/(2\sigma^2) = \\
&= [\mathrm{E}\,||\boldsymbol{y}^* - \boldsymbol{\mu}||^2 + \mathrm{E}\,||\hat{\boldsymbol{y}} - \boldsymbol{\mu}||^2 - 2\,\mathrm{E}\{(\boldsymbol{y}^* - \boldsymbol{\mu})^T(\hat{\boldsymbol{y}} - \boldsymbol{\mu}\}]/(2\sigma^2),
\end{aligned}
$$

as $||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2 = ||(\boldsymbol{y} - \boldsymbol{\mu}) - (\hat{\boldsymbol{y}} - \boldsymbol{\mu})||^2$. We go on as follows:

$$
\begin{aligned}
BC &= [\mathrm{E}\,||\boldsymbol{y}^* - \boldsymbol{\mu}||^2 + \mathrm{E}\,||\hat{\boldsymbol{y}} - \boldsymbol{\mu}||^2 - 2\,\mathrm{E}\{(\boldsymbol{y}^* - \boldsymbol{\mu})^T(\hat{\boldsymbol{y}} - \boldsymbol{\mu})\}]/(2\sigma^2) - \\
&\quad - [\mathrm{E}\,||\boldsymbol{y} - \boldsymbol{\mu}||^2 + \mathrm{E}\,||\hat{\boldsymbol{y}} - \boldsymbol{\mu}||^2 - 2\,\mathrm{E}\{(\boldsymbol{y} - \boldsymbol{\mu})^T(\hat{\boldsymbol{y}} - \boldsymbol{\mu})\}]/(2\sigma^2) = \\
&= \mathrm{E}\{(\boldsymbol{y} - \boldsymbol{\mu})^T(\hat{\boldsymbol{y}} - \boldsymbol{\mu})\}/(2\sigma^2),
\end{aligned}
$$

because conditionally on $\boldsymbol{b}$,

$$\mathrm{E}\,||\boldsymbol{y}^* - \boldsymbol{\mu}||^2 = \mathrm{E}\,||\boldsymbol{y} - \boldsymbol{\mu}||^2,$$

and

$$\mathrm{E}\{(\boldsymbol{y}^* - \boldsymbol{\mu})^T(\boldsymbol{y} - \boldsymbol{\mu})\} = \{\mathrm{E}(\boldsymbol{y}^* - \boldsymbol{\mu})\}^T\,\mathrm{E}(\boldsymbol{y} - \boldsymbol{\mu}) = 0.$$

Note that, conditionally on $\boldsymbol{b}$, $\mathrm{E}\{(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{\mu}\} = 0$, so that

$$BC = \mathrm{E}\{(\boldsymbol{y} - \boldsymbol{\mu})^T\hat{\boldsymbol{y}}/\sigma^2.$$

We substitute $\hat{\boldsymbol{y}}$ by $\boldsymbol{H}_1\boldsymbol{y}$, where $\boldsymbol{H}_1$ is the hat matrix and depends only on $\boldsymbol{X}, \boldsymbol{Z}$ and $\boldsymbol{W}$. This leads to

$$
\begin{aligned}
BC &= \mathrm{E}\{(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{H}_1\boldsymbol{y}\}/\sigma^2 = \\
&= \mathrm{E}\{(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{H}_1(\boldsymbol{y} - \boldsymbol{\mu})\}/\sigma^2 = \\
&= \mathrm{E}[\{\mathrm{tr}\{\boldsymbol{H}_1(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T\}]/\sigma^2 = \\
&= \mathrm{tr}[\boldsymbol{H}_1\,\mathrm{E}\{(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T\}]/\sigma^2 = \\
&= \mathrm{tr}\{\boldsymbol{H}_1\}.
\end{aligned}
$$

This is true because $\mathrm{E}\{(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T|\boldsymbol{b}\} = \mathrm{Cov}(\boldsymbol{y}|\boldsymbol{b}) = \sigma^2\boldsymbol{I}_n$.

Bias correction term for AIC in linear mixed-effects models with serial correlation

The deduction of the bias correction term BC for a linear mixed-effects model with serial correlation has some similarities to the one in the former subsection, however, the covariance matrix of the serial correlation has to be taken into account, which causes some changes. We now suppose that the residual variance σ^2, the covariance matrix of the random effects $\boldsymbol{W}$ as well as the variance of the serial correlation ξ^2 are known.

Let again $g(y|b)$ be the conditional likelihood of the model, and $\boldsymbol{y}^*$ be the prediction data set such that $\boldsymbol{y}^*$ and $\boldsymbol{y}$ are independent conditional on the true random effects. The number of measurements per individual n_i goes to infinity, so that we skip the index i. Apart from that, $\boldsymbol{y}_i = \boldsymbol{\mu}_i + \boldsymbol{D}_i + \boldsymbol{\epsilon}_i$ and $\boldsymbol{y}_i^* = \boldsymbol{\mu}_i + \boldsymbol{D}_i + \boldsymbol{\epsilon}_i^*$, as well as $\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$. Finally, $\boldsymbol{D}_i + \boldsymbol{\epsilon}_i, \boldsymbol{D_i} + \boldsymbol{\epsilon}_i^* \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i} + \xi^2\boldsymbol{\rho}(|\boldsymbol{u}_i|))$. For simplicity, we denote this covariance matrix by $\boldsymbol{\Sigma}_i$.

We can now write down the bias correction term as

$$\begin{aligned} BC &= \mathrm{E}_y \log g\{\boldsymbol{y}|\hat{\boldsymbol{b}}(\boldsymbol{y})\} - \mathrm{E}_y\,\mathrm{E}_y^* \log g\{\boldsymbol{y}^*|\hat{\boldsymbol{b}}(\boldsymbol{y})\} = \\ &= \mathrm{E}_y\{-\frac{1}{2}\sum_{i=1}^{I} n\log 2\pi + \log(\det(\boldsymbol{\Sigma}_i)) + (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)\} - \\ &- \mathrm{E}_y\,\mathrm{E}_y^*\{-\frac{1}{2}\sum_{i=1}^{I} n\log 2\pi + \log(\det(\boldsymbol{\Sigma}_i)) + (\boldsymbol{y}_i^* - \hat{\boldsymbol{y}}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i^* - \hat{\boldsymbol{y}}_i)\}. \end{aligned}$$

We now use the Cholesky decomposition and substitute $\boldsymbol{\Sigma}_i$ by $\boldsymbol{L}_i\boldsymbol{L}_i^T$. Thus,

$$(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = ||\boldsymbol{L}_i^T(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)||^2,$$

and we continue as follows:

$$\begin{aligned} BC &= \mathrm{E}_y\{-\frac{1}{2}In\log(2\pi) - \frac{1}{2}\sum_{i=1}^{I}\log(\det(\boldsymbol{\Sigma}_i)) - \frac{1}{2}\sum_{i=1}^{I}||\boldsymbol{L}_i^T(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)||^2\} - \\ &- \mathrm{E}_y\,\mathrm{E}_{y^*}\{-\frac{1}{2}In\log(2\pi) - \frac{1}{2}\sum_{i=1}^{I}\log(\det(\boldsymbol{\Sigma}_i)) - \frac{1}{2}\sum_{i=1}^{I}||\boldsymbol{L}_i^T(\boldsymbol{y}_i^* - \hat{\boldsymbol{y}}_i)||^2\} = \\ &= \mathrm{E}\{\frac{1}{2}\sum_{i=1}^{I}||\boldsymbol{L}_i^T(\boldsymbol{y}_i^* - \hat{\boldsymbol{y}}_i)||^2 - \frac{1}{2}\sum_{i=1}^{I}||\boldsymbol{L}_i^T(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)||^2\} = \\ &= \frac{1}{2}\sum_{i=1}^{I}[\mathrm{E}\,||\boldsymbol{L}_i^T\boldsymbol{y}_i^* - \boldsymbol{L}_i^T\boldsymbol{\mu}_i||^2 + \mathrm{E}\,||\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i||^2 - 2\,\mathrm{E}\{(\boldsymbol{L}_i^T\boldsymbol{y}_i^* - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)^T(\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)\}] - \\ &- \frac{1}{2}\sum_{i=1}^{I}[\mathrm{E}\,||\boldsymbol{L}_i^T\boldsymbol{y}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i||^2 + \mathrm{E}\,||\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i||^2 - 2\,\mathrm{E}\{(\boldsymbol{L}_i^T\boldsymbol{y}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)^T(\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)\}]. \end{aligned}$$

This is true because

$$||(\boldsymbol{L}_i^T\boldsymbol{y}_i - \boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i)||^2 = ||(\boldsymbol{L}_i^T\boldsymbol{y}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}) - (\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\boldsymbol{\mu})||^2,$$

which works analogously for $\boldsymbol{y}_i^*$. Note that, conditionally on $\boldsymbol{b}_i$,

$$\mathrm{E}\,||\boldsymbol{L}_i^T\boldsymbol{y}_i^* - \boldsymbol{L}_i^T\boldsymbol{\mu}_i||^2 = \mathrm{E}\,||\boldsymbol{L}_i^T\boldsymbol{y}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i||^2,$$

so that

$$BC = \sum_{i=1}^{I}[\mathrm{E}\{(\boldsymbol{L}_i^T\boldsymbol{y}_i - \boldsymbol{L}_i^T\boldsymbol{\mu})^T(\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T)\boldsymbol{\mu}\} - \mathrm{E}\{(\boldsymbol{L}_i^T\boldsymbol{y}_i^* - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)^T(\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)\}] =$$
$$= \sum_{i=1}^{I}\mathrm{E}\{(\boldsymbol{L}_i^T\boldsymbol{y}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)^T(\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)\},$$

because, conditionally on $\boldsymbol{b}$,

$$\mathrm{E}\{(\boldsymbol{L}_i^T\boldsymbol{y}_i^* - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)^T(\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\boldsymbol{\mu})\} = \{\mathrm{E}(\boldsymbol{L}_i^T\boldsymbol{y}_i^* - \boldsymbol{L}_i^T\boldsymbol{\mu}_i)\}^T\,\mathrm{E}(\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - \boldsymbol{L}_i^T\boldsymbol{\mu}_i) = 0.$$

We can now write

$$BC = \sum_{i=1}^{I}[\mathrm{E}\{(\boldsymbol{L}_i^T\boldsymbol{y}_i - \boldsymbol{L}_I^T\boldsymbol{\mu})^T\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - (\boldsymbol{L}_i^T\boldsymbol{y}_i - \boldsymbol{L}_I^T\boldsymbol{\mu})^T\boldsymbol{L}_i^T\boldsymbol{\mu}_i\}] =$$
$$= \sum_{i=1}^{I}[\mathrm{E}\{(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{L}_i\boldsymbol{L}_i^T\hat{\boldsymbol{y}}_i - (\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{L}_i\boldsymbol{L}_i^T\boldsymbol{\mu}_i\}] =$$
$$= \sum_{i=1}^{I}[\mathrm{E}\{(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\hat{\boldsymbol{y}}_i - (\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\}] =$$
$$= \sum_{i=1}^{I}[\mathrm{E}\{(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\hat{\boldsymbol{y}}_i\}].$$

This is true because, conditionally on $\boldsymbol{b}_i$,

$$\mathrm{E}\{(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\} = \mathrm{E}\{(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i = 0.$$

We now substitute $\hat{\boldsymbol{y}}_i$ by $\boldsymbol{H}_{1i}\boldsymbol{y}$, where $\boldsymbol{H}_{1i}$ stands for the rows of the hat matrix $\boldsymbol{H}_1$ concerning individual i. This leads to

$$BC = \sum_{i=1}^{I}[\mathrm{E}\{(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{H}_{1i}\boldsymbol{y}\}] =$$
$$= \sum_{i=1}^{I}[\mathrm{E}\{(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{H}_{1i}(\boldsymbol{y} - \boldsymbol{\mu})],$$

where we use that $-\mathrm{E}[(\boldsymbol{y}_1-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{H}_{1i}\boldsymbol{\mu}] = 0$. As the resulting expression is a scalar, we can write

$$\begin{aligned} BC &= \sum_{i=1}^{I}[\mathrm{E}\{\mathrm{tr}((\boldsymbol{y}_i-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{H}_{1i}(\boldsymbol{y}-\boldsymbol{\mu}))\}] = \\ &= \sum_{i=1}^{I}[\mathrm{E}\{\mathrm{tr}(\boldsymbol{H}_{1i}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1})\}] = \\ &= \sum_{i=1}^{I}[\mathrm{tr}\{\boldsymbol{H}_{1i}\,\mathrm{E}((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu}_i)^T)\boldsymbol{\Sigma}_i^{-1}\}]. \end{aligned}$$

We can partition the vector $(\boldsymbol{y}-\boldsymbol{\mu})$ in two parts, where the subscript i denotes all values concerning individual i and subscript $-i$ stands for all values except the ones of individual i:

$$(\boldsymbol{y}-\boldsymbol{\mu}) = \begin{pmatrix} \boldsymbol{y}_{-i}-\boldsymbol{\mu}_{-i} \\ \boldsymbol{y}_i-\boldsymbol{\mu}_i \end{pmatrix}$$

Thus we can continue as follows:

$$\begin{aligned} BC &= \sum_{i=1}^{I}[\mathrm{tr}\{\boldsymbol{H}_{1i}\,\mathrm{E}\left(\begin{pmatrix} \boldsymbol{y}_{-i}-\boldsymbol{\mu}_{-i} \\ \boldsymbol{y}_i-\boldsymbol{\mu}_i \end{pmatrix}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)^T\right)\boldsymbol{\Sigma}_i^{-1}\}] = \\ &= \sum_{i=1}^{I}[\mathrm{tr}\{\boldsymbol{H}_{1i}\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{\Sigma_i} \end{pmatrix}\boldsymbol{\Sigma}_i^{-1}\}]. \end{aligned}$$

This is valid because

$$\mathrm{E}\{(\boldsymbol{y}_i-\boldsymbol{\mu}_i)(\boldsymbol{y}_i-\boldsymbol{\mu}_i)^T|\boldsymbol{b}_i\} = \mathrm{Cov}(\boldsymbol{y}_i|\boldsymbol{b}_i) = \boldsymbol{\Sigma}_i$$

and because $(\boldsymbol{y}_{-i}-\boldsymbol{\mu}_{-i})$ and $(\boldsymbol{y}_i-\boldsymbol{\mu}_i)$ are conditionally independent given $\boldsymbol{b}$, leading to expected values of zero. In a final step, $\boldsymbol{H}_{1ii}$ stands for the block-diagonal parts of $\boldsymbol{H}_1$ that refer to individual i, and we can write the bias correction term as

$$\begin{aligned} BC &= \sum_{i=1}^{I}[\mathrm{tr}\{\boldsymbol{H}_{1i}\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{I} \end{pmatrix}\}] = \\ &= \sum_{i=1}^{I}[\mathrm{tr}\{\boldsymbol{H}_{1ii}\}] = \\ &= \mathrm{tr}\{\boldsymbol{H}_1\}. \end{aligned}$$

Web Appendix B

Hat matrix for linear mixed-effects models without serial correlation

The hat matrix for a linear mixed-effects model is deducted in Appendix 1 of Vaida and Blanchard (2005). This deduction works as follows:

The general linear-mixed-effects model without serial correlation of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$$

is rewritten, so that it is now of the form

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\delta} + \boldsymbol{e},$$

where

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{0} & -\boldsymbol{I} \end{pmatrix}, \text{ and } \boldsymbol{e} = \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{b} \end{pmatrix}.$$

The covariance matrix of the new residual vector $\boldsymbol{e}$ is now

$$\mathrm{Cov}(\boldsymbol{e}) = \mathrm{diag}(\sigma^2\boldsymbol{I}, \boldsymbol{W}) = \sigma^2\,\mathrm{diag}(\boldsymbol{I}, \boldsymbol{W}_0),$$

where $\boldsymbol{W}_0 = \sigma^{-2}\boldsymbol{W}$. Note that the matrix $\boldsymbol{W}_0$ is positive definite and can therefore be decomposed in $\boldsymbol{W}_0 = (\boldsymbol{\Delta}^T\boldsymbol{\Delta})^{-1}$.

Let $\Gamma = \mathrm{diag}(\boldsymbol{I}, \boldsymbol{\Delta})$ and multiply both sides in $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\delta} + \boldsymbol{e}$ with $\boldsymbol{\Gamma}$. This leads to

$$\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\delta} + \boldsymbol{w},$$

where

$$\boldsymbol{\Gamma}\boldsymbol{Y} = \boldsymbol{Y}, \quad \boldsymbol{M} = \boldsymbol{\Gamma}\boldsymbol{U} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{0} & -\boldsymbol{\Delta} \end{pmatrix} \text{ and } \boldsymbol{w} = \boldsymbol{\Gamma}\boldsymbol{e} = \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\Delta}\boldsymbol{b} \end{pmatrix}.$$

Now $\boldsymbol{w} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. This leads to the least-squares estimator of $\boldsymbol{\delta}$, which has the form

$$\hat{\boldsymbol{\delta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}^T & \boldsymbol{M} \end{pmatrix}^{-1} \boldsymbol{M}^T\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{M}^T & \boldsymbol{M} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{Z}^T \end{pmatrix} \boldsymbol{y}.$$

Looking at the fitted vector $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{b}} = \boldsymbol{H}_1\boldsymbol{y}$ leads to the hat matrix

$$\boldsymbol{H}_1 = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{M}^T & \boldsymbol{M} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{Z}^T \end{pmatrix}.$$

If only the trace of the hat matrix is needed, its calculation can be simplified by rearranging the parts of the hat matrix, as in Greven and Kneib (2010):

$$\mathrm{tr}\left\{ \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{W}_0^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} \end{pmatrix} \right\}.$$

Hat matrix for linear mixed-effects models with serial correlation and nugget effect

The deduction of the hat matrix for a linear mixed-effects model with serial correlation works relatively analogously to the one without serial correlation: the model is now of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{D} + \boldsymbol{\epsilon}.$$

this can again be rewritten to

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\delta} + \boldsymbol{e},$$

where $\boldsymbol{Y}, \boldsymbol{U}$ and $\boldsymbol{\delta}$ are defined as above, and

$$\boldsymbol{e} = \begin{pmatrix} \boldsymbol{\epsilon} + \boldsymbol{D} \\ \boldsymbol{b} \end{pmatrix}.$$

The covariance matrix of the new residual vector $\boldsymbol{e}$ is now

$$\mathrm{Cov}(\boldsymbol{e}) = \mathrm{diag}(\sigma^2\boldsymbol{I} + \xi^2\boldsymbol{R}, \boldsymbol{W}) = \sigma^2\,\mathrm{diag}(\boldsymbol{I} + \sigma^{-2}\xi^2\boldsymbol{R}, \boldsymbol{W}_0),$$

where $\boldsymbol{R} = \mathrm{diag}(\boldsymbol{\rho}(|\boldsymbol{u}_1|), ..., \boldsymbol{\rho}(|\boldsymbol{u}_I|))$ is the covariance matrix of $\boldsymbol{D}$ and $\boldsymbol{W}_0 = \sigma^{-2}\boldsymbol{W}$. The matrix $\boldsymbol{W}_0$ can again be decomposed in $\boldsymbol{W}_0 = (\boldsymbol{\Delta}^T\boldsymbol{\Delta})^{-1}$.

Let $\Gamma = \mathrm{diag}(\boldsymbol{I}, \boldsymbol{\Delta})$ and multiply both sides in $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\delta} + \boldsymbol{e}$ with $\boldsymbol{\Gamma}$. This leads to

$$\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\delta} + \boldsymbol{w},$$

where

$$\boldsymbol{\Gamma}\boldsymbol{Y} = \boldsymbol{Y}, \quad \boldsymbol{M} = \boldsymbol{\Gamma}\boldsymbol{U} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{0} & -\boldsymbol{\Delta} \end{pmatrix} \text{ and } \boldsymbol{w} = \boldsymbol{\Gamma}\boldsymbol{e} = \begin{pmatrix} \boldsymbol{\epsilon} + \boldsymbol{D} \\ \boldsymbol{\Delta}\boldsymbol{b} \end{pmatrix}.$$

Now $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ with $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \mathrm{diag}(\sigma^2\boldsymbol{I} + \xi^2\boldsymbol{R}, \sigma^2\boldsymbol{I})$. This leads to the weighted least-squares estimator of $\boldsymbol{\delta}$, which has the form

$$\hat{\boldsymbol{\delta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{pmatrix} = \left(\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{Y} = \left(\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{M}\right)^{-1}\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{Z}^T \end{pmatrix}(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1}\boldsymbol{y}.$$

Looking at the fitted vector $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{b}} = \boldsymbol{H}_1\boldsymbol{y}$ leads to the hat matrix

$$\boldsymbol{H}_1 = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{pmatrix}\left(\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{M}\right)^{-1}\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{Z}^T \end{pmatrix}(\sigma^2\boldsymbol{I} + \xi^2\boldsymbol{R})^{-1}.$$

As the calculation of this hat matrix can be quite time-consuming or even impossible for large data sets and rich models, it might be useful to rewrite

the matrix $\left(\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{M}\right)$ as follows:

$$\begin{aligned}
\left(\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{M}\right) &= \begin{pmatrix}\boldsymbol{X}^T & \boldsymbol{0}\\ \boldsymbol{Z}^T & -\boldsymbol{\Delta}^T\end{pmatrix}\begin{pmatrix}\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R} & \boldsymbol{0}\\ \boldsymbol{0} & \sigma^2\boldsymbol{I}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{X} & \boldsymbol{Z}\\ \boldsymbol{0} & -\boldsymbol{\Delta}\end{pmatrix}\\
&= \begin{pmatrix}\boldsymbol{X}^T & \boldsymbol{0}\\ \boldsymbol{Z}^T & -\boldsymbol{\Delta}^T\end{pmatrix}\begin{pmatrix}(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \sigma^{-2}\boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{X} & \boldsymbol{Z}\\ \boldsymbol{0} & -\boldsymbol{\Delta}\end{pmatrix}\\
&= \begin{pmatrix}\boldsymbol{X}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1} & \boldsymbol{0}\\ \boldsymbol{Z}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1} & -\sigma^{-2}\boldsymbol{\Delta}^T\boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{X} & \boldsymbol{Z}\\ \boldsymbol{0} & -\boldsymbol{\Delta}\end{pmatrix}\\
&= \begin{pmatrix}\boldsymbol{X}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1}\boldsymbol{X} & \boldsymbol{X}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1}\boldsymbol{X} & \boldsymbol{Z}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1}\boldsymbol{Z}+\sigma^{-2}\boldsymbol{\Delta}^T\boldsymbol{I}\boldsymbol{\Delta}\end{pmatrix}\\
&= \begin{pmatrix}\boldsymbol{X}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1}\boldsymbol{X} & \boldsymbol{X}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1}\boldsymbol{X} & \boldsymbol{Z}^T(\sigma^2\boldsymbol{I}+\xi^2\boldsymbol{R})^{-1}\boldsymbol{Z}+\boldsymbol{W}^{-1}\end{pmatrix}.
\end{aligned}$$

Hat matrix for linear mixed-effects models with serial correlation and no nugget effect

There are only a few minor changes to the deduction with nugget effect. In the case without nugget effect, $\boldsymbol{e}$ is reduced to

$$\boldsymbol{e} = \begin{pmatrix}\boldsymbol{D}\\ \boldsymbol{b}\end{pmatrix}.$$

The covariance matrix of the residual vector $\boldsymbol{e}$ is now

$$\mathrm{Cov}(\boldsymbol{e}) = \mathrm{diag}(\xi^2\boldsymbol{R}, \boldsymbol{W}).$$

We can now directly decompose $\boldsymbol{W}$, so that $\boldsymbol{W} = (\boldsymbol{\Delta}^T\boldsymbol{\Delta})^{-1}$.

Let $\Gamma = \mathrm{diag}(\boldsymbol{I}, \boldsymbol{\Delta})$ and multiply both sides in $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\delta} + \boldsymbol{e}$ with $\boldsymbol{\Gamma}$. This leads to

$$\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\delta} + \boldsymbol{w},$$

where

$$\boldsymbol{\Gamma}\boldsymbol{Y} = \boldsymbol{Y}, \quad \boldsymbol{M} = \boldsymbol{\Gamma}\boldsymbol{U} = \begin{pmatrix}\boldsymbol{X} & \boldsymbol{Z}\\ \boldsymbol{0} & -\boldsymbol{\Delta}\end{pmatrix} \text{ and } \boldsymbol{w} = \boldsymbol{\Gamma}\boldsymbol{e} = \begin{pmatrix}\boldsymbol{D}\\ \boldsymbol{\Delta}\boldsymbol{b}\end{pmatrix}.$$

Now $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ with $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \mathrm{diag}(\xi^2\boldsymbol{R}, \boldsymbol{I})$. This leads to the weighted least-squares estimator of $\boldsymbol{\delta}$, which has the form

$$\hat{\boldsymbol{\delta}} = \begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{b}}\end{pmatrix} = \left(\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{Y} = \left(\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{M}\right)^{-1}\begin{pmatrix}\boldsymbol{X}^T\\ \boldsymbol{Z}^T\end{pmatrix}(\xi^2\boldsymbol{R})^{-1}\boldsymbol{y}.$$

Looking at the fitted vector $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{b}} = \boldsymbol{H}_1\boldsymbol{y}$ leads to the hat matrix

$$\boldsymbol{H}_1 = \begin{pmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{pmatrix}\left(\boldsymbol{M}^T{\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1}\boldsymbol{M}\right)^{-1}\begin{pmatrix}\boldsymbol{X}^T\\ \boldsymbol{Z}^T\end{pmatrix}(\xi^2\boldsymbol{R})^{-1}.$$

Web Figure 1

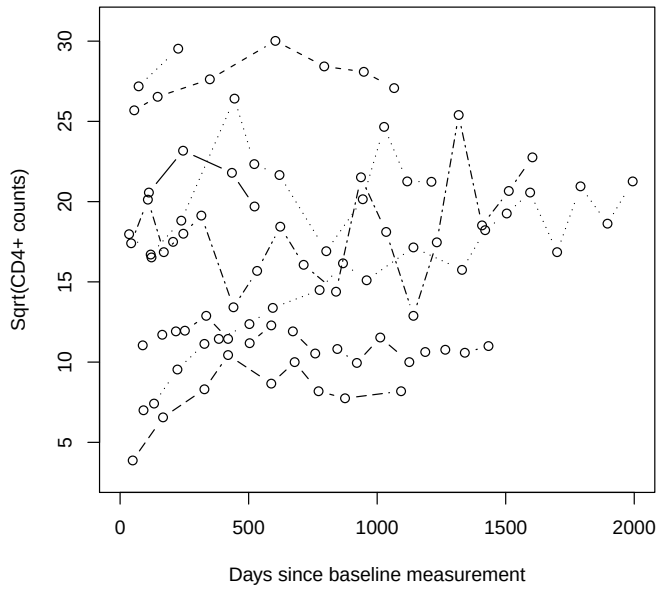

Figure 1: Exemplary course of the square root of the CD4+ counts

References

Greven, S. and Kneib, T. (2010). On the behavior of marginal and conditional AIC in linear mixed models. *Biometrika* **97,** 773–789.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92,** 351–370.

Supplementary material: Predictive cross-validation for choice of linear mixed-effects models with application to data from the Swiss HIV Cohort Study

Julia Braun, Leonhard Held and Bruno Ledergerber

March 24, 2011

1 General description

This is a short description of the `R` functions that can be used for performing univariate cross-validation based on the logarithmic or the continuous ranked probability score as described in the above named article. These programs were developed using `R`, version 2.10.1, on a laptop with Mac OS X, version 10.5.8, using a 2.4 GHz Intel Core 2 Duo processor.

The functions described in the following sections can be used for models with random intercept, random intercept and slope as well as models with just serial correlation with or without nugget effect or additional random effects, as can be fitted using the `lme` or `gls` commands from the `R` package `nlme`.

2 Univariate cross-validation

2.1 Function `cv.univ`

This function just needs the desired model and the data set and extracts the type of random effects and serial correlation used in the respective model. This information is then passed to the function `cross.scores` which performs the actual cross-validation (see next subsection) and returns the results.

Usage:

```
> cv.univ(model, dat)
```

Arguments:

- `model`: A fitted `gls` or `lme` model.

- `dat`: The data set used in the model fitting process. All measurements of the same individual should be in consecutive rows.

Value: A matrix containing the expected value and variance of the preditive distribution of each observation, as well as the associated logarithmic and continuous ranked probability score and the PIT and BOT values.

Functions used: `cross.scores`

2.2 Function `cross.scores`

This function is called by the function `cv.univ` and performs the actual cross-validation for any acceptable model. In each cross-validation step, the data set is brought into the right form, i.e. each observation is left out once. In the next step, the appropriate predictive distribution for this observation is calculated, and finally, the scores are obtained.

Usage:

```
> cross.scores(model, dat, type = "ri", cor.type = "Exp")
```

Arguments:

- `model`: A fitted `gls` or `lme` model.
- `dat`: The data set used in the model fitting process. All measurements of the same individual should be in consecutive rows.
- `type`: `"ri"` in the case of just a random intercept, `"ris"` for a model with random intercept AND random slope; default:`" ri"`.
- `cor.type`: `"Exp"` for exponential or `"Gauss"` for Gaussian correlation; default: `"Exp"`.
- `nugget`: `TRUE` if a nugget effect is included in the model, `FALSE` otherwise.

Value: A matrix containing the expected value and variance of the preditive distribution of each observation, as well as the associated logarithmic and continuous ranked probability score and the PIT and BOT values.

Functions used: `pred.cross.int`, `pred.cross.intslope`, `pred.cross.ser`, `pred.cross.int.ser`, `pred.cross.intslope.ser`, `pred.cross.ser.nonug`, `pred.cross.int.ser.nonug`, `pred.cross.intslope` `univ.scores`.

2.3 Function `pred.cross.int`

This is one of the different functions that can be called to calculate the univariate predictive distribution for a specific time point and combination of covariates. It is applicable for models with just a random intercept without any serial correlation.

Usage:

```
> pred.cross.int(y, obs, model, j, all, n = 1, nr, gr)
```

Arguments:

- `y`: Reduced data set for cross-validation, i.e. without the observation to be predicted.
- `obs`: Covariates for the observation to be predicted; must be brought into correct form using the function `model.matrix`.
- `model`: A fitted `gls` or `lme` model.
- `j`: Indicator for the individual from the original data set. If all individuals' IDs are stored in the vector `all` (see below) in the order of their appearance, `j` is the index of the respective individual.
- `all`: Vector with individual IDs.
- `n`: Number of predictions, which for univariate cross-validation always equals 1. Default: 1.
- `nr`: Column index of dependent variable.
- `gr`: Column index of grouping variable.

Value: Vector with expected value `E` and variance `Var` of the predictive distribution.

2.4 Function `pred.cross.intslope`

This is one of the different functions that can be called to calculate the univariate predictive distribution for a specific time point and combination of covariates. It is applicable for models with a random intercept and additional random slope without any serial correlation.

Usage:

```
> pred.cross.intslope(y, obs, model, j, all, n = 1, nr, gr)
```

Arguments:

- `y`: Reduced data set for cross-validation, i.e. without the observation to be predicted.
- `obs`: Covariates for the observation to be predicted; must be brought into correct form using the function `model.matrix`.
- `model`: A fitted `gls` or `lme` model.
- `j`: Indicator for the individual from the original data set. If all individuals' IDs are stored in the vector `all` (see below) in the order of their appearance, `j` is the index of the respective individual.
- `all`: Vector with individual IDs.
- `n`: Number of predictions, which for univariate cross-validation always equals 1. Default: 1.
- `nr`: Column index of dependent variable.
- `gr`: Column index of grouping variable.

Value: Vector with expected value `E` and variance `Var` of the predictive distribution.

2.5 Function `pred.cross.ser`

This is one of the different functions that can be called to calculate the univariate predictive distribution for a specific time point and combination of covariates. It is applicable for models with serial correlation and no random effects. This fuction is applicable for models including a nugget effect in the serial correlation structure.

Usage:

```
> pred.cross.ser(y, obs, model, j, all, cor.type = "Exp", n = 1, nr, gr,
+     tr)
```

Arguments:

- `y`: Reduced data set for cross-validation, i.e. without the observation to be predicted.
- `obs`: Covariates for the observation to be predicted; must be brought into correct form using the function `model.matrix`.
- `model`: A fitted `gls` or `lme` model.
- `j`: Indicator for the individual from the original data set. If all individuals' IDs are stored in the vector `all` (see below) in the order of their appearance, `j` is the index of the respective individual.

- `all`: Vector with individual IDs.
- `cor.type`: `"Exp"` for exponential or `"Gauss"` for Gaussian correlation; default: `"Exp"`.
- `n`: Number of predictions, which for univariate cross-validation always equals 1. Default: 1.
- `nr`: Column index of dependent variable.
- `gr`: Column index of grouping variable.
- `tr`: Column index of time variable.

Value: Vector with expected value `E` and variance `Var` of the predictive distribution.

2.6 Function `pred.cross.int.ser`

This is one of the different functions that can be called to calculate the univariate predictive distribution for a specific time point and combination of covariates. It is applicable for models with random intercept and serial correlation. This fuction is applicable for models including a nugget effect in the serial correlation structure.

Usage:

```
> pred.cross.int.ser(y, obs, model, j, all, cor.type = "Exp", n = 1, nr,
+     gr, tr)
```

Arguments:

- `y`: Reduced data set for cross-validation, i.e. without the observation to be predicted.
- `obs`: Covariates for the observation to be predicted; must be brought into correct form using the function `model.matrix`.
- `model`: A fitted `gls` or `lme` model.
- `j`: Indicator for the individual from the original data set. If all individuals' IDs are stored in the vector `all` (see below) in the order of their appearance, `j` is the index of the respective individual.
- `all`: Vector with individual IDs.
- `cor.type`: `"Exp"` for exponential or `"Gauss"` for Gaussian correlation; default: `"Exp"`.
- `n`: Number of predictions, which for univariate cross-validation always equals 1. Default: 1.

- `nr`: Column index of dependent variable.
- `gr`: Column index of grouping variable.
- `tr`: Column index of time variable.

Value: Vector with expected value `E` and variance `Var` of the predictive distribution.

2.7 Function `pred.cross.intslope.ser`

This is one of the different functions that can be called to calculate the univariate predictive distribution for a specific time point and combination of covariates. It is applicable for models with random intercept, slope and serial correlation. This fuction is applicable for models including a nugget effect in the serial correlation structure.

Usage:

```
> pred.cross.intslope.ser(y, obs, model, j, all, cor.type = "Exp", n = 1,
+     nr, gr, tr)
```

Arguments:

- `y`: Reduced data set for cross-validation, i.e. without the observation to be predicted.
- `obs`: Covariates for the observation to be predicted; must be brought into correct form using the function `model.matrix`.
- `model`: A fitted `gls` or `lme` model.
- `j`: Indicator for the individual from the original data set. If all individuals' IDs are stored in the vector `all` (see below) in the order of their appearance, `j` is the index of the respective individual.
- `all`: Vector with individual IDs.
- `cor.type`: `"Exp"` for exponential or `"Gauss"` for Gaussian correlation; default: `"Exp"`.
- `n`: Number of predictions, which for univariate cross-validation always equals 1. Default: 1.
- `nr`: Column index of dependent variable.
- `gr`: Column index of grouping variable.
- `tr`: Column index of time variable.

Value: Vector with expected value `E` and variance `Var` of the predictive distribution.

2.8 Function `pred.cross.ser.nonug`

This is one of the different functions that can be called to calculate the univariate predictive distribution for a specific time point and combination of covariates. It is applicable for models with serial correlation and no random effects. This fuction is applicable for models without a nugget effect in the serial correlation structure.

Usage:

```
> pred.cross.ser.nonug(y, obs, model, j, all, cor.type = "Exp", n = 1, nr,
+     gr, tr)
```

Arguments:

- `y`: Reduced data set for cross-validation, i.e. without the observation to be predicted.
- `obs`: Covariates for the observation to be predicted; must be brought into correct form using the function `model.matrix`.
- `model`: A fitted `gls` or `lme` model.
- `j`: Indicator for the individual from the original data set. If all individuals' IDs are stored in the vector `all` (see below) in the order of their appearance, `j` is the index of the respective individual.
- `all`: Vector with individual IDs.
- `cor.type`: `"Exp"` for exponential or `"Gauss"` for Gaussian correlation; default: `"Exp"`.
- `n`: Number of predictions, which for univariate cross-validation always equals 1. Default: 1.
- `nr`: Column index of dependent variable.
- `gr`: Column index of grouping variable.
- `tr`: Column index of time variable.

Value: Vector with expected value `E` and variance `Var` of the predictive distribution.

2.9 Function `pred.cross.int.ser.nonug`

This is one of the different functions that can be called to calculate the univariate predictive distribution for a specific time point and combination of covariates. It is applicable for models with random intercept and serial correlation. This fuction is applicable for models without a nugget effect in the serial correlation structure.

Usage:

```
> pred.cross.int.ser.nonug(y, obs, model, j, all, cor.type = "Exp", n = 1,
+     nr, gr, tr)
```

Arguments:

- `y`: Reduced data set for cross-validation, i.e. without the observation to be predicted.
- `obs`: Covariates for the observation to be predicted; must be brought into correct form using the function `model.matrix`.
- `model`: A fitted `gls` or `lme` model.
- `j`: Indicator for the individual from the original data set. If all individuals' IDs are stored in the vector `all` (see below) in the order of their appearance, `j` is the index of the respective individual.
- `all`: Vector with individual IDs.
- `cor.type`: `"Exp"` for exponential or `"Gauss"` for Gaussian correlation; default: `"Exp"`.
- `n`: Number of predictions, which for univariate cross-validation always equals 1. Default: 1.
- `nr`: Column index of dependent variable.
- `gr`: Column index of grouping variable.
- `tr`: Column index of time variable.

Value: Vector with expected value `E` and variance `Var` of the predictive distribution.

2.10 Function `pred.cross.intslope.ser.nonug`

This is one of the different functions that can be called to calculate the univariate predictive distribution for a specific time point and combination of covariates. It is applicable for models with random intercept, slope and serial correlation. This fuction is applicable for models without a nugget effect in the serial correlation structure.

Usage:

```
> pred.cross.intslope.ser.nonug(y, obs, model, j, all, cor.type = "Exp",
+     n = 1, nr, gr, tr)
```

Arguments:

- `y`: Reduced data set for cross-validation, i.e. without the observation to be predicted.

- `obs`: Covariates for the observation to be predicted; must be brought into correct form using the function `model.matrix`.
- `model`: A fitted `gls` or `lme` model.
- `j`: Indicator for the individual from the original data set. If all individuals' IDs are stored in the vector `all` (see below) in the order of their appearance, `j` is the index of the respective individual.
- `all`: Vector with individual IDs.
- `cor.type`: `"Exp"` for exponential or `"Gauss"` for Gaussian correlation; default: `"Exp"`.
- `n`: Number of predictions, which for univariate cross-validation always equals 1. Default: 1.
- `nr`: Column index of dependent variable.
- `gr`: Column index of grouping variable.
- `tr`: Column index of time variable.

Value: Vector with expected value `E` and variance `Var` of the predictive distribution.

2.11 Function `univ.scores`

This function calculates the univariate logarithmic score, continuous ranked probability score, PIT and BOT values from a normal predictive distribution and the true observation.

Usage:

```
> univ.scores(y.obs, E, sd)
```

Arguments:

- `y.obs`: True observed value.
- `E`: Expected value of predictive distribution.
- `sd`: Standard deviation of predictive distribution.

Value: Vector with LogS, CRPS, PIT and BOT values.

PAPER II

Choice of generalized linear mixed models using predictive cross-validation

Julia Braun, Daniel Sabanés Bové & Leonhard Held

Paper under review for *Computational Statistics and Data Analysis*.

Choice of generalized linear mixed models using predictive crossvalidation

Julia Braun, Daniel Sabanés Bové, and Leonhard Held

Division of Biostatistics, Institute for Social and Preventive Medicine,

University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland

The choice of generalized linear mixed models is difficult, because it involves the selection of both fixed and random effects. Classical criteria like Akaike's information criterion (AIC) are often not suitable for the latter task, and others which are useful in linear mixed models are difficult to extend to the generalized case, especially for overdispersed data. We suggest to use a predictive leave-one-out crossvalidation approach that can be applied for choosing both fixed and random effects, also in models with overdispersion, and is based on proper scoring rules. An attractive feature of this approach is the fact that the model has to be fitted just once to the data set, which makes computations fast and convenient. As the calculation of the leave-one-out predictive distribution is not possible analytically, we show how an iteratively weighted least squares algorithm combined with some analytic approximations can be used for this task. Two applications of the methodology to binary and count data are demonstrated, and comparisons with two other methods are provided.

Keywords: Predictive model choice; proper scoring rules; Poisson regression; logistic regression; conditional AIC; overdispersion.

1 Introduction

Model choice in linear or generalized linear models is a relatively easy task, and much information on various criteria and techniques is available. If, however, these models are extended to contain random effects to accommodate e.g. longitudinal data, choosing a model becomes much more challenging. One reason for this is the fact that in addition to the selection of covariates, a decision on the kind and number of random effects has to be made. Classical criteria like Akaike's information criterion (AIC, Akaike, 1973) or the Bayesian information criterion (BIC, see e.g. Schwarz, 1978 or Claeskens and Hjort, 2008) are not sufficient for this task and must be adapted.

Before applying any criterion for model choice, a decision on the focus of the desired analysis has to be made, namely if the main interest lies on the fixed effects (population level) or if information on the random effects (individual or cluster level) is desired. In the case of linear mixed models, Pauler (1998) addresses the choice of fixed effects using a BIC version suitable for unbalanced longitudinal data. For choosing random effects, Pauler et al. (1999) use a boundary Laplace approximation to obtain a BIC version including an additional term for boundary correction.

In terms of the AIC, the consequences of the main focus of inference are highlighted in Vaida and Blanchard (2005), where the authors show that the generally known, classical version of the AIC – the marginal AIC – should only be applied for the selection of fixed effects. For deciding on the inclusion of random effects, the conditional AIC (cAIC) is introduced, for which the effective degrees of freedom needed in the penalty term can be calculated as in Hodges and Sargent (2001). This concept is extended to deliver more reliable results by Liang et al. (2008), but the numerical calculation is quite involved and time-consuming, as shown by Greven and Kneib (2010).

Unfortunately, all these concepts only relate to linear mixed models and are difficult to extend to generalized linear mixed models. A Bayesian approach to the simultaneous selection of fixed and random effects via zero-inflated (truncated) normal priors on fixed effects and on elements of the decomposed random effects covariance matrix is presented by Cai and Dunson (2006). Jiang et al. (2008) and Nguyen and Jiang (2012) suggest to use fence methods for the selection of the fixed effects in generalized linear mixed models. These methods choose models from a range of candidate models by setting and subsequently restricting boundaries of some suitable criterion (see also Claeskens and Hjort (2008, p. 273)). An analytic deduction of the cAIC is impossible, as shown by Donohue et al. (2011), but the authors suggest an asymptotic approximation which includes the effective degrees of freedom as defined by Lu et al. (2007). They note, however, that their asymptotic approximation may not be reliable in certain settings and propose using bootstrap methods instead. An asymptotically unbiased estimator of the cAIC for use with generalized linear mixed models is presented by Yu and Yau (2012), which seems to be quite similar to the above mentioned approximation by Donohue et al. (2011). Lian (2012) proposes another unbiased estimator of the cAIC to be used for Poisson regression models, which involves a high number of model fits and might thus be unsuitable for large data sets.

In this article, we introduce an alternative approach to selecting generalized linear mixed models for longitudinal data from a predictive point of view. By using mean crossvalidated proper scores (see Gneiting and Raftery, 2007) as criterion for model choice, both fixed and random effects can be selected leading to a model with the best predictive abilities. The crossvalidated logarithmic score is closely related to the AIC in linear models (see Stone, 1977 or Pawitan, 2001) and to the cAIC in linear mixed models (Braun et al., 2012), so that its application in the case of generalized linear mixed models seems promising.

Other than in the linear mixed model, the (leave-one-out) predictive distribution that is necessary for the calculation of the proper scores cannot be deducted analytically. To solve this problem, we propose to use an iteratively weighted least squares (IWLS)

algorithm with prior distribution as described in Gamerman (1997), complemented by some analytic approximations. To shorten the computation time, we reduce the number of necessary model fits to just one, using "mixed" crossvalidation as described by Marshall and Spiegelhalter (2003). This approach is by far less time-consuming than full leave-one-out crossvalidation, and Braun et al. (2012) have shown empirically that the results from both crossvalidation approaches are comparable for the linear mixed model.

This article is organized as follows: We review the basics of generalized linear mixed models in Section 2 and show the proper scoring rules needed for comparing predictive distributions in Section 3. The predictive crossvalidation approach based on a Bayesian IWLS algorithm is presented and outlined specifically for the cases of logistic regression, Poisson regression, and a Poisson model including overdispersion, in Section 4. Applications to binary and count data are discussed in Section 5, followed by a comparison with two other approximate estimators of the cAIC. Section 6 adds a summary and some general discussion.

2 Generalized linear mixed models

Generalized linear mixed models for longitudinal data are generally defined as follows (see for example Fahrmeir and Tutz, 2001): Assume that each individual $i = 1, \dots, I$ provides observations y_{ij} at time points t_j with $j = 1, \dots, J$. For simplicity, we assume that the time points are the same for each individual, but this is not a necessary precondition. Let the vectors $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ contain covariates relating to the fixed and random effects, respectively, then the linear predictor is defined as

$$\eta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i,$$

with p fixed effects $\boldsymbol{\beta}$ and q random effects $\boldsymbol{b}_i$. It is assumed that the random effects are normally distributed $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{Q})$ and that, conditional on the random effects $\boldsymbol{b}_i$, different measurements of the same subject are independent. Their conditional expected value $\mu_{ij} = \mathrm{E}(y_{ij} \,|\, \boldsymbol{b}_i)$ is related to the linear predictor via an appropriate link function g, so that

$$g(\mu_{ij}) = \eta_{ij}.$$

Given the fixed and random effects, the conditional distribution of the response y_{ij} belongs to an exponential family with density

$$f(y_{ij} \,|\, \boldsymbol{b}_i) = \exp\Big\{ \frac{y_{ij}\theta_{ij} - \kappa(\theta_{ij})}{\phi} + c(y_{ij}, \phi) \Big\},$$

where θ_{ij} is the natural parameter, ϕ is a dispersion parameter and κ and c are functions that depend on the type of exponential family. Note that $\mu_{ij} = \kappa'(\theta_{ij})$, where κ' denotes the first derivative of κ.

The two non-Gaussian regression models that are applied most often in this context are binary logistic and log-linear Poisson regression models. In the case of logistic regression,

each observation y_{ij} has a Bernoulli distribution with probability

$$p_{ij} = P\{y_{ij} = 1\} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}. \quad (1)$$

The log-linear Poisson model assumes the expectation

$$\lambda_{ij} = \exp(\eta_{ij}).$$

Overdispersion can be included in a model by estimating an additional random effect for each observation, as explained e.g. by Collett (2003, p. 293). This offers the possibility to account for additional variability without having to assume that there is a linear relationship over time of each individual's measurements (as is the case if a random slope is included for each individual). The mixed Poisson model with overdispersion has a linear predictor of the form

$$\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + a_{ij},$$

where $a_{ij} \sim N(0, \xi^2)$ represents the additional random effect for each observation y_{ij}. Note that $\boldsymbol{b}_i$ and a_{ij} are mutually independent. Information on fitting and interpreting generalized linear mixed models as well as other suitable models for discrete longitudinal data can e.g. be found in Molenberghs and Verbeke (2005).

3 Proper scoring rules

As stated in the introduction, the selection of random effects in generalized linear mixed models based on any selection criterion that is currently available is difficult. Therefore, we suggest to use proper scoring rules as a criterion for model choice instead. They are a simple, yet effective instrument for assessing predicive distributions. This way, both the choice of fixed and random effects is possible. In this paper, all scores are positively oriented, so that a larger score denotes a model with better predictive abilities. By taking into account not only the distance between a point prediction and the true value, but also the predictive variance, proper scores cover both sharpness, i.e. the concentration of a predictive distribution, and calibration, i.e. the consistency between the predictive distribution and the actual observations. An essential property of a scoring rule is propriety, which ensures that both calibration and sharpness are addressed simultaneously (Winkler, 1996) and is defined as follows: The expected value of a proper score is maximal if the observation is a realization from the presumed predictive distribution. General information on the concept of proper scoring rules as well as examples for the case of continuous predictive distributions can e.g. be found in Gneiting and Raftery (2007) and Gneiting et al. (2007).

A well-known proper scoring rule for binary predictions is the Brier score (Brier, 1950). It is also called quadratic score and is defined as

$$\mathrm{BS}(Y, y_{\mathrm{obs}}) = -(p_Y - y_{\mathrm{obs}})^2,$$

where p_Y stands for the predicted probabiliy of the outcome and $y_{\text{obs}} \in \{0, 1\}$ is the actual observation.

The proper logarithmic score (LS) is well suited to assess the predictive abilities of any regression model if the density f_Y of the predictive distribution Y is known. It is defined as the value of the log density of Y at the actually observed value y_{obs}:

$$\text{LS}(Y, y_{\text{obs}}) = \log f_Y(y_{\text{obs}}). \tag{2}$$

The LS is sometimes criticized for being a so-called local score, because it only evaluates the predictive density at y_{obs}, but ignores all other values of $f_Y(y)$. However, it has proved to be a valuable tool for model criticism, and its calculation is simple once the density of the predictive distribution is known.

An alternative to the LS is the Dawid-Sebastiani score (DSS), which was proposed by Dawid and Sebastiani (1999) and is used as preditive model choice criterion e.g. by Held, Rufibach and Balabdaoui (2010). It is defined as

$$\text{DSS}(Y, y_{\text{obs}}) = -\frac{1}{2}\Big\{\log(\sigma_Y^2) + \Big(\frac{y_{\text{obs}} - \mu_Y}{\sigma_Y}\Big)^2\Big\} \tag{3}$$

and has the advantage that only the first two central moments μ_Y and σ_Y^2 of the predictive distribution of Y are necessary for its calculation. This is particularly important in our crossvalidation approach which is presented later on. Alternative model assessment tools for use with models for count data are given by Czado et al. (2009).

4 Predictive crossvalidation

Conducting a full crossvalidation often turns out to be very time-consuming and in some cases even impossible due to the size of the respective data set and the complexity of the model. As a potential alternative, Marshall and Spiegelhalter (2003) present mixed predictive model checks, where the model is fitted just once to the complete data set. In each step of the following crossvalidation, the estimated individual random effects and the concrete observation are ignored, and a forecast is generated based on the estimated fixed effects and the hyperparameters of the random effects. As the omitted observation influences the random effects only via their estimated covariance matrix, but not directly, the introduced conservatism is only moderate, and thus tolerable. This approach has been used before to select linear mixed models by Braun et al. (2012), and in different contexts among others by Riebler and Held (2010) and Held, Schrödle and Rue (2010).

In order to conduct this predictive crossvalidation approach for mixed Poisson models, each of the competing models is fitted only once to the whole data set. After fitting the model, one observation y_{ij} from the data set is left out, and the predictive distribution for this observation is calculated based on the remaining observations $\boldsymbol{y}_{i,-j}$. Note again that only the estimated fixed effects parameters $\hat{\boldsymbol{\beta}}$ and the estimated covariance of the random effects $\hat{\boldsymbol{Q}}$ are used for this task, but not the initially estimated individual random effects parameters. From this the proper scores described in Section 3 can be calculated.

This procedure is then repeated for each observation y_{ij}. For comparison, we show the predicted expected value of the linear predictor obtained using full crossvalidation:

$$\mathrm{E}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_{-ij} + \boldsymbol{z}_{ij}^T \hat{\boldsymbol{b}}_{i,-j},$$

where all measurements except y_{ij} are used for the estimation of both $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$. In contrast, these coefficients are replaced by their estimates $\hat{\boldsymbol{\beta}}$ from the initial model fit if our proposed crossvalidation approach is applied.

4.1 Bayesian iteratively weighted least squares algorithm

Several steps are necessary to obtain the leave-one-out predictive distribution for the observation y_{ij}. First, estimates of the conditional expectation $\mathrm{E}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j})$ and covariance matrix $\mathrm{Cov}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j})$ are needed. To do this, an algorithm suggested by West (1985) and further elaborated in Gamerman (1997) is used: Treat $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ as a given offset, and $\boldsymbol{z}_{ij}$ like "normal" covariates, so that a Bayesian estimation of the "regression coefficients" $\boldsymbol{b}_i$ can be performed. Combining the likelihood with the prior distribution $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \hat{\boldsymbol{Q}})$ leads to the approximate posterior distribution

$$\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j} \overset{a}{\sim} N(\boldsymbol{m}_{ij}, \boldsymbol{C}_{ij}),$$

whose parameters are obtained using the following Bayesian iteratively weighted least squares (IWLS) algorithm. Note that this algorithm is equivalent to a so-called penalized iteratively reweighted least squares (PIRLS) algorithm (Bates and DebRoy, 2004) and is also used in the `R` package `lme4` for the estimation of random effects in generalized linear mixed models (see Bates, 2012). It works as follows: After choosing some starting values for $\boldsymbol{m}_{ij}^{(0)}$, for which we use the estimated random effects $\hat{\boldsymbol{b}}_i$ from the model fit, the estimates $\boldsymbol{m}_{ij}^{(k)}$ and $\boldsymbol{C}_{ij}^{(k)}$ in the kth iteration are

$$\boldsymbol{m}_{ij}^{(k)} = \boldsymbol{C}_{ij}^{(k)} \boldsymbol{z}_{i,-j} \boldsymbol{W}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)}) \tilde{\boldsymbol{y}}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)}).$$

and

$$\boldsymbol{C}_{ij}^{(k)} = \{\hat{\boldsymbol{Q}}^{-1} + \boldsymbol{z}_{i,-j} \boldsymbol{W}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)}) \boldsymbol{z}_{i,-j}^T\}^{-1}$$

The "design matrix" $\boldsymbol{z}_{i,-j}$ of dimension $q \times (J-1)$ contains data from all time points of individual i except the time point of interest t_j. The elements of the response vector $\tilde{\boldsymbol{y}}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)})$ are the pseudo observations

$$\tilde{y}_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)} + \{y_{is} - \mu_{is}(\boldsymbol{m}_{ij}^{(k-1)})\} g'\{\mu_{is}(\boldsymbol{m}_{ij}^{(k-1)})\} \tag{4}$$

for $s \neq j$, where $\mu_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = g^{-1}(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)})$. As described in Gamerman (1997), the weights $W_{is}(\boldsymbol{m}_{ij}^{(k-1)})$ are defined via

$$W_{is}^{-1}(\boldsymbol{m}_{ij}^{(k-1)}) = \kappa''(\theta_{is}(\boldsymbol{m}_{ij}^{(k-1)}))\{g'(\mu_{is}(\boldsymbol{m}_{ij}^{(k-1)}))\}^2, \tag{5}$$

so that the matrix containing all weights is $\boldsymbol{W}_{i,-j}(\boldsymbol{m}_{ij}^{(k-1)}) = \text{diag}\{W_{is}(\boldsymbol{m}_{ij}^{(k-1)})\}_{s \neq j}$.

These iterations are terminated as soon as

$$\max\left\{\frac{|\boldsymbol{m}_{ij}^{(k)} - \boldsymbol{m}_{ij}^{(k-1)}|}{|\boldsymbol{m}_{ij}^{(k-1)}|}\right\} < \epsilon,$$

with e.g. $\epsilon = 10^{-6}$, where $|\cdot|$ and max are taken over all components of $\boldsymbol{m}_{ij}^{(k)}$ and $\boldsymbol{m}_{ij}^{(k-1)}$.

The second step after having obtained $\text{E}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) \approx \boldsymbol{m}_{ij}$ and $\text{Cov}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) \approx \boldsymbol{C}_{ij}$ involves the calculation of the first two moments of the predictive distribution. This depends on the specific regression model and can either be done by approximation or numerical integration. The required steps for this calculation as well as the specific formulae for (4) and (5) are discussed below for the case of binary logistic and log-linear Poisson regression with and without overdispersion. A short discussion of its applicability in further generalized linear mixed models can be found in Section 6.

4.2 Predictive crossvalidation for mixed logistic regression models

If applied to a mixed logistic regression model, formulae (4) and (5) have the form

$$\tilde{y}_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)} + \frac{y_{is} \cdot (1 + \exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}))^2}{\exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)})} - \exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}) - 1$$

and

$$W_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \frac{\exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)})}{(1 + \exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}))^2},$$

leading to estimates $\text{E}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) \approx \boldsymbol{m}_{ij}$ and $\text{Cov}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) \approx \boldsymbol{C}_{ij}$. The expected value and variance of η_{ij} are then

$$\text{E}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T \boldsymbol{m}_{ij} =: \tau \tag{6}$$

and

$$\text{Var}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \boldsymbol{z}_{ij}^T \boldsymbol{C}_{ij} \boldsymbol{z}_{ij} =: \sigma^2. \tag{7}$$

In order to obtain the predictive probability $P\{y_{ij} = 1 \,|\, \boldsymbol{y}_{i,-j}\}$, the mixed logistic regression model is rewritten as a latent variable model. Expression (1) corresponds to

$$\omega_{ij} = \eta_{ij} + \epsilon_{ij},$$

where $y_{ij} = 1$ if $\omega_{ij} \geq 0$, $y_{ij} = 0$ if $\omega_{ij} < 0$ and ϵ_{ij} follows a standard logistic distribution. This can be approximated by a normal distribution (Zeger et al., 1988), so that

$$\epsilon_{ij} \overset{a}{\sim} N(0, c)$$

with $c = (15/16)^2 \cdot \pi^2/3$. Thus,

$$\omega_{ij} \overset{a}{\sim} N(\tau, \sigma^2 + c),$$

so that $P\{y_{ij} = 1 \,|\, \boldsymbol{y}_{i,-j}\} = \int_0^\infty N(x \,|\, \tau, \sigma^2 + c)dx$ can be calculated based on the distribution function of the normal distribution and subsequently used for the calculation of the BS and the LS.

4.3 Predictive crossvalidation for mixed Poisson regression models

In the case of a log-linear Poisson regression model, the pseudo observations $\tilde{y}_{is}(\boldsymbol{m}_{ij}^{(k-1)})$ for $s \neq j$ are

$$\tilde{y}_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)} + y_{is} \exp(-\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} - \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}) - 1, \tag{8}$$

and the weights $W_{is}(\boldsymbol{m}_{ij}^{(k-1)})$ are

$$W_{is}(\boldsymbol{m}_{ij}^{(k-1)}) = \exp(\boldsymbol{x}_{is}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{is}^T \boldsymbol{m}_{ij}^{(k-1)}). \tag{9}$$

The calculation of the first two moments of the predictive distribution is straightforward. Expected value τ and variance σ^2 of η_{ij} are obtained as in (6) and (7). As η_{ij} is (approximately) normally distributed, $\exp(\eta_{ij}) = \lambda_{ij}$ is log-normally distributed with

$$\mathrm{E}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \exp\Big(\tau + \frac{1}{2}\sigma^2\Big) \tag{10}$$

and

$$\mathrm{Var}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \{\exp(\sigma^2) - 1\} \exp(2\tau + \sigma^2). \tag{11}$$

In a final step, the predictive expectation and variance of the observation y_{ij} are

$$\mathrm{E}(y_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \mathrm{E}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) = \exp\Big(\tau + \frac{1}{2}\sigma^2\Big)$$

and

$$\begin{aligned}
\mathrm{Var}(y_{ij} \,|\, \boldsymbol{y}_{i,-j}) &= \mathrm{E}\{\mathrm{Var}(y_{ij} \,|\, \eta_{ij}, \boldsymbol{y}_{i,-j})\} + \mathrm{Var}\{\mathrm{E}(y_{ij} \,|\, \eta_{ij}, \boldsymbol{y}_{i,-j})\} \\
&= \mathrm{E}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) + \mathrm{Var}(\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}) \\
&= \exp\Big(\tau + \frac{1}{2}\sigma^2\Big) + \{\exp(\sigma^2) - 1\} \exp(2\tau + \sigma^2).
\end{aligned}$$

These two values allow the calculation of the DSS (3) for each observation from the data set and its respective prediction, and subsequently of the mean DSS. To obtain the LS, however, the predictive expectation and variance are not sufficient, because the density of the predictive distribution has to be known. This problem can be solved using two distinct approaches: The first possibility is an approximation of the log-normal distribution of $\lambda_{ij} \,|\, \boldsymbol{y}_{i,-j}$ via the gamma distribution, which is performed by matching the first two moments of these two distributions:

The two parameters of the gamma distribution can be chosen in such a way that its expected value and variance equal (10) and (11), respectively. This is the case for a gamma distribution $G(\alpha, \phi)$ with density

$$f(\lambda_{ij}) = \frac{1}{\Gamma(\alpha)} \Big(\frac{\alpha}{\phi}\Big)^{\alpha} \lambda_{ij}^{\alpha-1} \exp\Big(-\frac{\lambda_{ij}\alpha}{\phi}\Big)$$

and parameters

$$\alpha = \frac{\mathrm{E}(\lambda_{ij})^2}{\mathrm{Var}(\lambda_{ij})} = \frac{1}{\exp(\sigma^2) - 1}$$

and

$$\phi = \mathrm{E}(\lambda_{ij}) = \exp\Big(\tau + \frac{1}{2}\sigma^2\Big).$$

Note that this approach works only if $\alpha > 1$, because it must be made sure that the respective gamma distribution has a mode larger than 0.

With λ_{ij} being approximately gamma distributed, the marginal distribution of y_{ij} follows a negative binomial distribution with density

$$\begin{aligned} f(y_{ij}) &= \int_0^\infty f(y_{ij} \,|\, \lambda_{ij}) f(\lambda_{ij}) d\lambda_{ij} \\ &= \frac{\Gamma(\alpha + y_{ij})}{\Gamma(\alpha)\Gamma(y_{ij} + 1)} \Big(\frac{\alpha}{\phi + \alpha}\Big)^\alpha \Big(\frac{\phi}{\phi + \alpha}\Big)^{y_{ij}}, \end{aligned}$$

as explained in Winkelmann (2008, p. 35). Evaluating this (log) density with mean parameter ϕ and size parameter α at the actual observation yields the desired LS. To ensure that the used approximation steps work reasonably well, we recommend to compare the resulting mean LS with the mean DSS.

Alternatively, one can simply use numerical integration, which should be reasonably quick if the data set is not too large. In that case, the density of the predictive distribution at the actual observation y_{obs} is obtained by the integral

$$f(y_{\mathrm{obs}}) = \int_0^\infty f(y_{\mathrm{obs}} \,|\, \lambda_{ij}) f(\lambda_{ij}) d\lambda_{ij}, \tag{12}$$

where λ_{ij} follows a log-normal distribution with parameters (10) and (11).

4.4 Predictive crossvalidation for mixed Poisson regression models with overdispersion

The predictive crossvalidation procedure with overdispersion works almost as in the ordinary Poisson case (without overdispersion), there are just some minor changes: Let

$$\boldsymbol{\eta}_i = \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{r}_i \boldsymbol{d}_i$$

be the model for all observations of individual i, where

$$\boldsymbol{d}_i = \begin{pmatrix} \boldsymbol{b}_i \\ \boldsymbol{a}_i \end{pmatrix}$$

is the vector containing all random effects of individual i, the design matrix of the fixed effects is

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{x}_{i1}^T \\ \vdots \\ \boldsymbol{x}_{iJ}^T \end{pmatrix}$$

and the design matrix for the random effects has the form

$$\boldsymbol{r}_i = \begin{pmatrix} \boldsymbol{z}_{i1}^T & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ \boldsymbol{z}_{iJ}^T & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

For the estimation of the conditional expected value of $\boldsymbol{d}_i \,|\, \boldsymbol{y}_{i,-j}$, we use the prior distribution $\boldsymbol{d}_{ij} \sim N(\boldsymbol{0}, \mathrm{diag}(\hat{\boldsymbol{Q}}, \hat{\xi}^2, \ldots, \hat{\xi}^2))$, and $\boldsymbol{r}_{i,-j}$ is $\boldsymbol{r}_i$ without the jth row and has to be used instead of $\boldsymbol{z}_{i,-j}$ in formulae (8) and (9). Apart from that, all remaining formulae from the IWLS algorithm stay the same. Note that the calculation of the expected value and variance of η_{ij} as in formulae (6) and (7) are now

$$\begin{aligned} \mathrm{E}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) &= \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}} + \boldsymbol{r}_{ij}^T \,\mathrm{E}(\boldsymbol{d}_i \,|\, \boldsymbol{y}_{i,-j}) \\ &= \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T \,\mathrm{E}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) + 0 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(\eta_{ij} \,|\, \boldsymbol{y}_{i,-j}) &= \boldsymbol{r}_{ij}^T \,\mathrm{Cov}(\boldsymbol{d}_i \,|\, \boldsymbol{y}_{i,-j}) \boldsymbol{r}_{ij} \\ &= \boldsymbol{z}_{ij}^T \,\mathrm{Cov}(\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}) \boldsymbol{z}_{ij} + \hat{\xi}^2, \end{aligned}$$

meaning that the expected value and variance of $a_{ij} \,|\, \boldsymbol{y}_{i,-j}$ are the moments of its prior distribution, and only the predictive expectation and variance of $\boldsymbol{b}_i \,|\, \boldsymbol{y}_{i,-j}$ need to be calculated.

5 Application

We illustrate the methods presented above using two well-known data sets from the literature.

5.1 Case study I: Xerophthalmia and respiratory disease in Indonesian children

The first data set is presented by Diggle et al. (2002, p. 4) and contains binary data on infections of the respiratory tract and xerophthalmia (dryness of the eye due to vitamin A deficiency) in Indonesian children, along with additional information on the children's age and height. Up to 6 measurements per child were collected in quarterly visits. 22 of the 275 patients were removed because they contributed only one measurement, so that the remaining data set consists of 1178 measurements of 253 children with 105 events of respiratory infection.

The variables sex, height in relation to age (percentage obtained from the United States National Center for health statistics), and presence of xerophthalmia are included in all models. As suggested by Diggle et al. (2002, p. 156 and 182), a model M1 with

age and age squared was compared to a model M2 with follow-up time and follow-up time squared instead. In the latter case, age at baseline and age squared at baseline are included as additional covariates. In addition, both models were fitted with two covariates representing an annual sine and cosine (denoted by "with season"). All models either include just a random intercept (RI) or, alternatively, an additional random slope (RIS) which depends on age (M1) or time (M2).

The mean BS and LS of these eight models are given in Table 1. Both scores clearly prefer models including a seasonal sine and cosine to the ones without that seasonal component, and the models with the lowest BS or, respectively, LS both use follow-up time instead of age. The BS chooses a model with just random intercept (BS: −0.07537), whereas the LS selects the model with additional random slope (LS: −0.27376).

Table 1: Respiratory infection: Mean BS and LS for the eight different models

	BS	LS
M1 (age), RI	−0.07732	−0.28124
M1 (age), RIS	−0.07695	−0.27917
M1 with season, RI	−0.07593	−0.27622
M1 with season, RIS	−0.07557	−0.27435
M2 (time), RI	−0.07644	−0.27678
M2 (time), RIS	−0.07665	−0.27680
M2 with season, RI	−0.07537	−0.27386
M2 with season, RIS	−0.07567	−0.27376

5.2 Case study II: Seizure counts

The second example analyses a frequently used data set used by Thall and Vail (1990), which was also shown in Diggle et al. (2002, p. 10), with counts of epileptic seizures as outcome. In this randomized crossover study patients were treated against partial epileptic seizures with either progabide (an anti-epileptic drug) or placebo and followed over four subsequent clinic visits. At each visit, they reported the number of epileptic seizures during the last two weeks. In our analyses, only the clinic visits before crossing over to the alternative treatment are analysed. Note that one patient was left out due to very unusual measurements, as suggested by Diggle et al. (2002, p. 164).

The data set contains information of 58 patients with four clinic visits each. The covariates used in all models are baseline seizure rate, treatment as well as a baseline-treatment interaction term, and the logarithm of age, completed by either the respective visit number or its square root serving as time variable. The following models are compared: One model with just a random intercept and a second model comprising an additional random slope, either using the visit number or its square root in both the fixed and the random effects part. In a second step, we add an indicator for the fourth visit to each of these models in order to account for markedly low counts at the last visit of each patient.

The first and second column of Table 2 display the mean crossvalidated DSS and LS of the eight competing models. We used the approximation of the log-normal distribution by a gamma distribution as explained in Section 4.3. For comparison, we also used numerical integration as in formula (12) and calculated the same integral with the approximate gamma distribution for λ_{ij}. All three methods lead to practically the same results. Note that in this example, the DSS and the LS put the models in exactly the same order. For both versions of the time variable, the model with random intercept and slope is preferred, and both scores show a clear preference for models including the indicator of the fourth visit. Finally, using the square root of time instead of the original time variable leads to an additional improvement, especially concerning the models with indicator of the fourth visit. All this shows that the model best suited for prediction is the model with random intercept and slope, based on the square root of the time variable and including a variable indicating the fourth visit, having a mean DSS of -1.9785 and a mean LS of -2.7620.

The third and fourth column of Table 2 show the mean scores for models with random intercept that include a random effect for each observation in order to incorporate overdispersion (denoted by "OD"). Fitting models with random intercept, slope and the random effect for each observation leads to overfitting and causes the variance of the random slope to be very small and the correlation between random slope and intercept to be 1. From this we can conclude that including an additional random slope does not ameliorate the overdispersion model.

Concerning the models with random intercept only, we can see that accounting for overdispersion leads to a remarkable improvement in both mean scores. In contrast, the differences between the four models are small, showing that all four models are equally useful for making predictions. Both scores again select models that include the indicator for the fourth visit, and the LS chooses the model with visit as time variable (LS: -2.5584), whereas the DSS slightly prefers the model with the square root of visit number (DSS: -1.756).

Table 2: Epileptic seizures: Mean DSS and LS for the eight different models with and without overdispersion (OD). "−" indicates that scores could not be calculated due to a singular covariance matrix.

	DSS	LS	DSS, OD	LS, OD
visit, RI	−2.0453	−2.7946	−1.7572	−2.5585
visit, RIS	−2.0262	−2.7881	−	−
sqrt(visit), RI	−2.0488	−2.7961	−1.7592	−2.5591
sqrt(visit), RIS	−2.0147	−2.7797	−	−
visit, RI, visit 4	−2.0153	−2.7866	−1.7566	−2.5584
visit, RIS, visit 4	−1.9953	−2.7749	−	−
sqrt(visit), RI, visit 4	−2.0152	−2.7866	−1.7560	−2.5585
sqrt(visit), RIS, visit 4	−1.9785	−2.7620	−	−

5.3 Comparison with other methods

We compare the results of our proposed method with the ones obtained using two other suggestions. The first alternative method is presented by Donohue et al. (2011) and involves an asymptotic version of the cAIC. The second method we compare our results to is the corrected conditional AIC (ccAIC) derived by Yu and Yau (2012), which can also be applied if the covariance matrix of the random effects is unknown. The authors kindly provided us with their Matlab programs which we translated into `R` functions and extended to the case of binary logistic regression. Unfortunately, these programs are so far only useable for models with just a random intercept and no random slope, for which reason the ccAIC could not be calculated for all candidate models in our applications. These two alternative methods can so far not be used for a Poisson model including an additional random effect to cover overdispersion. Note that both criteria seem to be very similar, which is also confirmed by the results in our applications.

Table 3: Logistic regression: Comparison with other methods; (c)cAIC values transformed by $-\frac{1}{2n}$

	LS	cAIC Donohue	ccAIC Yu
M1 (age), RI	−0.28124	−0.28254	−0.28254
M1 (age), RIS	−0.27917	−0.27868	–
M1 with season, RI	−0.27622	−0.27903	−0.27903
M1 with season, RIS	−0.27435	−0.27504	–
M2 (time), RI	−0.27678	−0.27954	−0.27955
M2 (time), RIS	−0.27680	−0.27802	–
M2 with season, RI	−0.27386	−0.27786	−0.27787
M2 with season, RIS	−0.27376	−0.27585	–

Table 4: Poisson regression: Comparison with other methods; (c)cAIC values transformed by $-\frac{1}{2n}$

	LS	cAIC Donohue	ccAIC Yu
visit, RI	−2.7946	−2.6686	−2.6687
visit, RIS	−2.7881	−2.6100	–
sqrt(visit), RI	−2.7961	−2.6698	−2.6699
sqrt(visit), RIS	−2.7797	−2.6022	–
visit, RI, visit 4	−2.7866	−2.6672	−2.6673
visit, RIS, visit 4	−2.7749	−2.6032	–
sqrt(visit), RI, visit 4	−2.7866	−2.6672	−2.6673
sqrt(visit), RIS, visit 4	−2.7620	−2.5931	–

Tables 3 and 4 again show the mean LS obtained with our proposed procedure, along with the cAIC by Donohue et al. (2011) (denoted by "Donohue") and the ccAIC by Yu

and Yau (2012) (denoted by "Yu", only for models with random intercept). To ease the comparisons, we put the cAIC and ccAIC values on the same scale as the mean LS by dividing them by $-2\sum J_i$, as has been done before in Braun et al. (2012).

Concerning the logistic regression model in Table 3, the results from the three different procedures differ only in the third decimal place. The models with just a random intercept are arranged in the same order by all three methods. If all models are taken into account (i.e. a random slope as well), the best three models and the worst model are clearly identified by our method and the cAIC, however, the overall order is slightly different and another model is chosen to be the best.

If we have a look at Table 4, we can see that the differences between our proposed method and the other two methods are larger than for logistic regression, but the results are still of similar magnitude. The ordering of the competing models is not identical but similar, and especially the decision which model is the best or worst is the same using all three methods. In order to find an explanation for the differences between mean LS and cAIC and ccAIC, we conducted some additional simulation studies. We found that if the true distribution of the data is Poisson and a Poisson model is fitted, there are only very small differences between the different methods. If, however, the data are overdispersed - as it is the case in our application - and come from a negative binomial distribution, the differences increase along with the amount of overdispersion.

To illustrate the comparison between the transferred mean LS and the cAIC, Figure 1 shows the respective values from both methods in both applications. We don't show the ccAIC, because there are no visible differences between cAIC and ccAIC. We can see that in both cases, the models with just random intercept perform worse than models with additional random slope. The order of the models with random intercept is equal, but small differences occur when a random slope is included. Summing up, it can be said that our proposed crossvalidation approach leads to results that are comparable to the other two possible methods, but not exactly the same, especially for count data with overdispersion.

6 Discussion

This paper has presented a novel predictive crossvalidation approach to model selection in generalized linear mixed models. The crossvalidated LS and BS or, respectively, DSS form a useful means for selecting both fixed and random effects. As the model has to be fitted only once, this approach is by far less time-consuming than a true leave-one-out crossvalidation.

We have demonstrated the calculation of mean proper scoring rules with our crossvalidation approach for the two most common generalized linear mixed models, i.e. logistic regression and Poisson regression (including overdispersion), but the approach is applicable more widely, often using procedures that are very similar to the ones used above. For example, overdispersed binomial data can be analyzed by adding an additional random effect for each observation to the linear predictor, and the predictive distribution is then obtained by using the formulation as latent variable model from Section 4.2. The

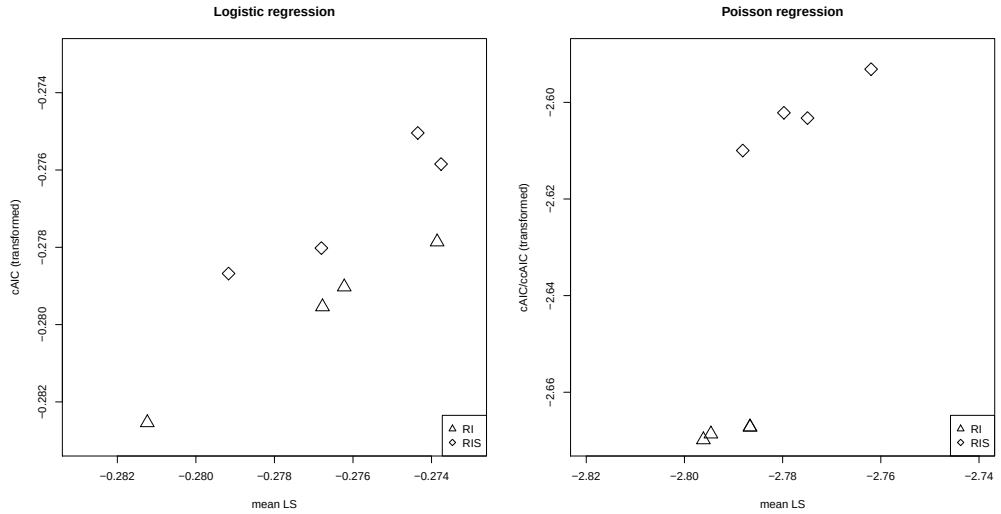

Figure 1: Comparison of transformed cAIC with mean LS for models with random intercept (RI) and models with additional random slope (RIS). Note that the values of two models with random intercept in the right plot are equal, for which reason only three RI models are visible.

first two moments of the predictive distribution in log-linear gamma regression models can be calculated analogously to Poisson regression models, and numerical intergation as in (12) can be used for the calculation of the LS. The derivation of the predictive distribution or its moments could be slightly more complicated in some other, more rarely used generalized linear mixed models, but it should be feasible in most cases.

The application of the IWLS algorithm in the Poisson case with overdispersion can cause problems in data sets with a large number of observations, because inversions of large matrices are needed. However, in our predictive crossvalidation approach it is only applied for one individual in the data set at a time, so that this should not be problematic. Note that Gamerman (1997) provides an alternative algorithm based on building blocks of correlated parameters, which can be used for larger data sets if needed.

Concerning the calculation of the predictive density for obtaining the LS for a mixed Poisson model, the approximation of the log-normal distribution via a gamma distribution can be realized in different ways. The matching of moments which we have applied could be problematic for certain forms of the respective distributions. As a possible alternative, we have tried minimizing the Kullback-Leibler distance between the two distributions, but this is much more time-consuming and often not applicable due to numerical problems. For this reason, it is advisable to calculate both the LS and the DSS in the case of a mixed Poisson model and see if the results are comparable.

In comparison with other possible approaches to model choice in generalized linear mixed models, our method has two decisive advantages: First, it can take into account overdispersion, which proves very useful in routine applications. Second, other approaches involve the multiplication and inversion of large matrices. This is not a problem in small data sets, but as soon as the data set gets large, the necessary calculations take considerable time or may not be possible at all. Our method can also deal with very large data sets, which makes it practical and useful in a wide set of situations.

References

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle, *in* B. N. Petrov and F. Csaki (eds), *International Symposium on Information Theory*, Budapest: Akademia Kiado, pp. 267–281.

Bates, D. (2012). Linear mixed model implementation in `lme4`, `http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf`.

Bates, D. and DebRoy, S. (2004). Linear mixed models and penalized least squares, *Journal of Multivariate Analysis* **91**(1): 1–17.

Braun, J., Held, L. and Ledergerber, B. (2012). Predictive cross-validation for the choice of linear mixed-effects models with application to data from the Swiss HIV Cohort Study, *Biometrics* **68**(1): 53–61.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability, *Monthly Weather Review* **78**(1): 1–3.

Cai, B. and Dunson, D. (2006). Bayesian covariance selection in generalized linear mixed models, *Biometrics* **62**(2): 446–457.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*, Cambridge University Press, Cambridge, Great Britain.

Collett, D. (2003). *Modelling Binary Data*, 2nd edn, Chapman & Hall/CRC.

Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data, *Biometrics* **65**(4): 1254–1261.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design, *The Annals of Statistics* **27**(1): 65–81.

Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd edn, Oxford University Press.

Donohue, M., Overholser, R., Xu, R. and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models, *Biometrika* **98**(3): 685–700.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, second edn, Springer.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models, *Statistics and Computing* **7**: 57–68. 10.1023/A:1018509429360.
URL: *http://dx.doi.org/10.1023/A:1018509429360*

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society, Ser. B* **69**: 243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation, *Journal of the American Statistical Association* **102**: 359–378.

Greven, S. and Kneib, T. (2010). On the behavior of marginal and conditional AIC in linear mixed models, *Biometrika* **97**(4): 773–789.

Held, L., Rufibach, K. and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions, *Biometrics* **66**(4): 1295–1305.

Held, L., Schrödle, B. and Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA., *in* T. Kneib and G. Tutz (eds), *Statistical Modelling and Regression Structures - Festschrift in Honour of Ludwig Fahrmeir*, Physica-Verlag, Heidelberg, Germany, pp. 91–110.

Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models, *Biometrika* **88**(2): 367–379.

Jiang, J., Rao, J., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection, *The Annals of Statistics* **36**(4): 1669–1692.

Lian, H. (2012). A note on conditional Akaike information for Poisson regression with random effects, *Electronic Journal of Statistics* **6**: 1–9.

Liang, H., Wu, H. and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models, *Biometrika* **95**(3): 773–778.

Lu, H., Hodges, J. S. and Carlin, B. P. (2007). Measuring the complexity of generalized linear hierarchical models, *The Canadian Journal of Statistics* **35**(1): 69–87.

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models, *Statistics in Medicine* **22**(10): 1649–1660.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer.

Nguyen, T. and Jiang, J. (2012). Restricted fence method for covariate selection in longitudinal data analysis, *Biostatistics* **13**(2): 303–314.

Pauler, D. (1998). The Schwarz criterion and related methods for normal linear models, *Biometrika* **85**(1): 13–27.

Pauler, D., Wakefield, J. and Kass, R. (1999). Bayes factors and approximations for variance component models, *Journal of the American Statistical Association* **94**(448): 1242–1253.

Pawitan, Y. (2001). *In All Likelihood - Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford, Great Britain.

Riebler, A. and Held, L. (2010). The analysis of heterogeneous time trends in multivariate age-period-cohort models, *Biostatistics* **11**(1): 57–69.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**(2): 461–464.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Ser. B* **39**(1): 44–47.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics* **46**(3): 657–671.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models, *Biometrika* **92**(2): 351–370.

West, M. (1985). Generalized linear models: Scale parameters, outlier accommodation, scale parameters and prior distributions, *in* J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (eds), *Bayesian Statistics 2*, Elsevier Science Publishers B. V., North Holland, pp. 531–558.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*, Springer.

Winkler, R. (1996). Scoring rules and the evaluation of probabilities, *Test* **5**: 1–60.

Yu, D. and Yau, K. K. W. (2012). Conditional Akaike information criterion for generalized linear mixed models, *Computational Statistics and Data Analysis* **56**(3): 629–644.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach, *Biometrics* **44**(4): 1049–1060.
URL: *http://www.jstor.org/stable/2531734*

PAPER III

Accounting for baseline differences and measurement error in the analysis of change over time

Julia Braun, Leonhard Held, Bruno Ledergerber & the Swiss HIV Cohort Study

Second revised version under review for *Statistics in Medicine*.

Accounting for baseline differences and measurement error in the analysis of change over time

Julia Braun*, Leonhard Held*, Bruno Ledergerber**,
and the Swiss HIV Cohort Study***

*Division of Biostatistics, Institute for Social and Preventive Medicine,

University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland

**Division of Infectious Diseases and Hospital Epidemiology,

University Hospital Zurich, University of Zurich, Rämistrasse 100, 8091 Zurich, Switzerland

If change over time is compared in several groups, it is important to take into account baseline values, so that the comparison is carried out under the same preconditions. As the observed baseline measurements are distorted by measurement error, it may not be sufficient to include them as covariate. By fitting a longitudinal mixed-effects model to all data including the baseline observations and subsequently calculating the expected change conditional on the underlying baseline value, a solution to this problem has been provided recently so that groups with the same baseline characteristics can be compared. In this article, we present an extended approach where a broader set of models can be used. Specifically, it is possible to include any desired set of interactions between the time variable and the other covariates, and also time-dependent covariates can be included. Additionally, the method is extended to adjust for baseline measurement error of other time-varying covariates. The methodology is applied to data from the Swiss HIV Cohort Study to address the question if a joint infection with HIV-1 and hepatitis C virus (HCV) leads to a slower increase of CD4+ lymphocyte counts over time after start of antiretroviral therapy.

Keywords: Longitudinal mixed-effects models; BIC; underlying baseline measurement; measurement error;

***The members of the Swiss HIV Cohort Study are: Aubert V, Barth J, Battegay M, Bernasconi E, Böni J, Bucher HC, Burton-Jeangros C, Calmy A, Cavassini M, Egger M, Elzi L, Fehr J, Fellay J,

1 Introduction

Longitudinal data reports observations from the same individual over time, thus gathering information on the temporal development of the variable of interest. If a comparison of change over time in different groups is of interest, special care has to be taken to adjust for potentially differring baseline measurements. When two or more groups are to be compared with respect to their development over time, this comparison should ideally relate to the same baseline value. This problem is further complicated by the fact that the true underlying baseline value can not be observed directly due to additional measurement error.

There are well-known techniques to analyse changes from baseline measurement when there is only one follow-up measurement, for example change scores and analysis of covariance (ANCOVA), see Vickers and Altman (2001) for a basic introduction. The effects of measurement error in change models and a possible correction method are investigated by Yanez III et al. (1998). Adjusting for measurement error may be particularly necessary if ANCOVA shall be performed using data from observational studies (Walter et al., 2011).

Unfortunately, literature and solutions for the case of several follow-up measurements are sparse, and almost no method can be found to deal with the case where the time points of the measurements are different for each individual. Treating the observed baseline measurement as explanatory variable is a quite popular, but insufficient technique, because the associated measurement error leads to biased results (Chambless and Davis, 2003) and larger standard errors (Verbeke et al., 2006). An ANCOVA-like technique for more than one follow-up measurement is described in Fitzmaurice et al. (2004, Section 5.6), but this method is only appropriate in randomized trials, not in observational studies. An overview of alternatives is given in Fitzmaurice et al. (2004, Section 5.7), however, only some of these methods can be used in observational studies, and none of them allows the comparison of groups based on the same underlying true baseline value. Other concepts to incorporate the measurement error in change models are proposed by Chambless and Roeback (1993) and Chambless and Davis (2003), but they are only applicable for balanced data and rely on additional information on the covariate variances from preliminary studies which might not be available in some cases.

An elegant and relatively easy solution to this problem that can also be used with unbalanced data is provided in Harrison et al. (2009), where the expected change from baseline, conditional on the true underlying baseline value, is obtained after fitting a linear mixed-effects model to all data. This technique delivers adjusted coefficients with respect to change from baseline, and the expected change can be obtained for any hy-

Francioli P, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Gorgievski M, Günthard H (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hirschel B, Hösli I, Kahlert C, Kaiser L, Keiser O, Kind C, Klimkait T, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Metzner K, Müller N, Nadal D, Pantaleo G, Rauch A (Chairman of the Scientific Board), Regenass S, Rickenbach M (Head of Data Center), Rudin C (Chairman of the Mother & Child Substudy), Schmid P, Schultze D, Schöni-Affolter F, Schüpbach J, Speck R, Taffé P, Tarr P, Telenti A, Trkola A, Vernazza P, Weber R, Yerly S.

pothetical value of the underlying true baseline measurement.

This method, however, is in its original form only useable for a limited set of model types: Only time-constant covariates can be included so far, and interaction terms between the time variable and all other covariates have to be included, leading to an unnecessarily distended model which can be hard to interpret. Moreover, if not only the outcome variable, but also a covariate is affected by measurement error, this can so far not be considered, so that the results may be strongly biased despite the use of the correction method. Luckily though, it can be extended to account for these additional tasks, which will be explained in this article.

It is organized as follows: In Section 2, we present a data set from the Swiss HIV Cohort Study (SHCS), choose a suitable model and explain why our proposed extended methodology is needed in this context. In Section 3, several correction methods to incorporate the underlying baseline measurements are presented and subsequently applied in Section 4, where, additionally, other models and the influence of the choice of covariate values are examined empirically. Finally, a summary and some discussion is added in Section 5.

2 Influence of HCV coinfection on HIV-1 disease progression

A joint infection with human immunodeficiency virus (HIV-1) and hepatitis C virus (HCV) is very common among intravenous drug users. HCV coinfection could possibly lead to different patterns of clinical progression of HIV-1 infection, however, the existence and nature of such an influence is still controversial. Many studies on that subject have been conducted which, especially in the last decade, have focussed on patients receiving highly active antiretroviral therapy (HAART), defined as regimens including at least three drugs, of which at least one non-nucleoside reverse-transcriptase inhibitor and/or a protease inhibitor.

These studies have used different approaches to model the influence of HCV coinfection on HIV-1 progression, among which survival techniques and longitudinal modelling are the most prominent. A Cox proportional hazards model was used by Greub et al. (2000) with events defined as progression to a new AIDS-defining opportunistic illness or death, as well as increase in CD4 cell counts of more than 50 cells/μl. While this study found a clear difference between HCV-seropositive and seronegative patients with respect to clinical progression and increase in CD4 cell counts, no evidence for such a relationship was reported in Rockstroh et al. (2005). Other studies applied similar methods to assess the effect of an HCV coinfection for patient collectives outside of Europe, e.g. Robbins et al. (2007) in the United States or Mbougua et al. (2010) in Cameroon.

The latter study additionally used linear mixed-effects models for the course of the square root of CD4 counts, but the baseline CD4 values were not taken into account, whereas Cheng et al. (2007) did include the baseline measurements and found a significant influence of HCV seropositivity on the CD4 cell count (not transformed), but only in patients adherent to their therapy. An overall significant difference in CD4 cell count was found by Egger et al. (2009). Similar methods were also used to search for other

variables that might influence the development of the CD4 cell counts, such as alcohol consumption in Samet et al. (2007). A different approach to assessing the influence of HCV seropositivity on the course of HIV-1 infection using longitudinal data is to focus on annual change in CD4 cell counts instead of the actual count, as has been done in Peters et al. (2009), where no effect of HCV coinfection was found.

2.1 Data set

Our data set obtained from the SHCS has been split into three parts: One contains data from patients that started potent antiretroviral therapy between June 1, 1996 and May 31, 1999, the early days of HAART therapy. At that time, the therapy options were limited, all being much more toxic and less efficatious than today. The second data set contains more recent data from patients that started potent antiretroviral therapy after May 31, 1999 and before December 31, 2004. This time period can be called a transition phase, where new drug classes were developed and less toxic combinations of drugs were introduced. The last data set covers the time after December 31, 2004, where more potent and efficacious treatments had been established, and even less toxic therapies with once-daily regimens lead to more quality of life for HIV patients. Looking at three data sets from these different time periods allows to examine if the effect of an HCV coinfection follows the same pattern when different therapy options dominate. In order to distinguish the three data sets from each other, we will denote the first data set by "O" (standing for "old"), the second data set by "M" ("middle") and the data set containing the most recent data with "N" ("new"). Only measurements within four years after therapy start are taken into account.

After deleting observations with missing covariate information, the data set "O" consists of 33196 measurements (including baseline measurements) from 2415 patients, of whom 1651 had a negative hepatitis C status, 134 showed an inactive, and 630 an active HCV coinfection, defined as detectable HCV RNA in plasma. The data set "M" includes 20471 measurements from 1490 patients, among them 1135 with negative, 75 with inactive and 280 with active hepatitis C. Finally, the last data set "N" covers 19570 measurements from 2030 patients, with 1737 HCV negative patients, 69 patients with inactive and 224 patients with active HCV coinfection.

The data sets contain potential covariates that could influence the course of the CD4 cell counts. They cover a wide range of information, e.g. on the patients' general health and social situation as well as detailed information on the HIV-1 infection and therapy. The mean baseline CD4 counts for the different groups range from 13.68 to 16.39.

In Figure 1 we see a moving average (bandwidth 4 months) of the change in the square root of CD4 cell counts for the groups with different hepatitis C states compared to the group without hepatitis C. We can see that in the middle and new data set, the curves representing active and inactive hepatitis C are quite similar. In the old data set, however, the inactive curve is not only closer to, but even slightly higher than the negative curve. One would expect that patients with inactive hepatitis C recover more quickly than patients with active hepatitis C and are comparable to patients with negative HCV status, which is not visible in neither of the curves. It has to be examined

if this still is the case after adjusting for relevant covariates.

2.2 Model construction

For unbalanced data sets like ours, that comprise a large number of measurements per subject at distinct time points, a linear mixed-effects model seems to be particularly useful (Diggle et al., 2002, p. 54 and p. 81). By including an individual random intercept and random slope in the model, it is possible to take into account stochastic variation between individuals. Moreover, this is a very flexible class of models that can be extended in many directions depending on the demands of the data. General information on linear mixed-effects models, their properties and construction can be found, among others, in Verbeke and Molenberghs (2000), Frees (2004) or Pinheiro and Bates (2004).

Two different random slope models were used to assess differences in the slopes of patients with and without HCV coinfection: In a first step, only patients with negative hepatitis C status were compared to patients who had had contact with the virus (i.e. combining patients from the active and inactive hepatitis C category) to clarify if there is any impact of HCV coinfection. After that, we had a look at a variable comprising all three hepatitis C categories separately, so that a possibly different influence of active and inactive HCV infection can be explored. The reference category was "negative" in all cases.

Based on information from HIV-1 experts concerning relevant influential variables, the following additional time-constant covariates were included in all models: AIDS at baseline, duration of use of nucleoside-analogue reverse-transcriptase inhibitors (NRTI) before potent antiretroviral therapy (in years), sex and age. To account for group differences in change over time, it was necessary to include an interaction term of the time variable with hepatitis C status in the model. Time was measured in years after start of HAART. Although many details of the chosen model were specified in advance, there remained several open questions which made the use of an appropriate criterion for model choice necessary. In particular, other interactions with time besides the one already included may lead to an improved model.

Experts also wished to include the decimal logarithm $\log_{10}$ of HIV-1 RNA in the models, because HIV-1 RNA is a central parameter for evaluating the strength of the HI virus. It would have to be included as a time-dependent covariate. In cases where HIV-1 RNA was so low that it was not measurable any more, we assumed it to be half of the respective detection limit. Thus, one problem is the non-negligible amount of measurement error in HIV-1 RNA, especially in measurements under the detection limit. As the main focus of our approach is to adjust for measurement error in the outcome variable, the inclusion of additional measurement error in a covariate is unsatisfactory. Moreover, HIV-1 RNA may be on the causal pathway between HCV status and CD4 count, so the interpretation of HIV-1 RNA as a covariate is difficult.

For these reasons, HIV-1 RNA was omitted in the mixed-effects models for CD4 cell counts. To address the problems explained above, we additionally use a bivariate model, where both the square root of CD4 counts and $\log_{10}$ of HIV-1 RNA are included as outcome variables and the covariates remain the same as in the univariate model.

2.3 Model choice

The Akaike information criterion (AIC) as well as the Bayesian information criterion (BIC) are frequently used to select the most suitable model from all candidate models. They have the advantage that they are in many cases easily calculated and provide a clear statement as to which model of a number of competing models should be chosen. Increasing model complexity is penalized more strongly when the BIC is applied, for which reason we decided to use this criterion as our primary basis for model choice. The model with the lowest BIC value of all candidates is considered to be most adequate.

The BIC is in general defined as

$$\mathrm{BIC} = -2\log L + p\log(n),$$

where L stands for the likelihood of the specific model, p is the number of estimated parameters, and n the number of measurements in the data set. However, in the case of linear mixed-effects models, this definition has to be modified (Pauler, 1998):

$$\mathrm{BIC} = -2\log L + \sum_{k=1}^{p_0} \log(n_k). \tag{1}$$

Here p_0 is the length of the coefficient vector $\boldsymbol{\beta}$, and the value of n_k depends on whether or not the k-th covariate has an associated random effect. If there is a random effect, n_k equals the number of individuals I, whereas it equals the number of measurements $\sum_{i=1}^{I} N_i$ if the covariate is only included as a fixed effect. Note that this version of the BIC is appropriate if only the choice of covariates without associated random effects is desired. If a decision on the inclusion of a random effect or model choice for individual predictions is intended, the so-called conditional AIC should be used. More information on this topic can be found in Vaida and Blanchard (2005) and Braun et al. (2012). Problems associated with the use of the BIC if a decision on the inclusion of a random effect is needed are discussed in Pauler et al. (1999).

The BIC version (1) for the use with longitudinal data was applied to decide on a model as suitable as possible for our purpose. Models with time or, alternatively, the square root of time were compared. In all cases, using the latter version of the time variable leads to considerably better models, which is consistent with the course of the CD4 cell counts seen in the descriptive analyses.

Additionally, (1) was used to decide how many interactions with the square root of time should be included in the model. We fitted 16 candidate models containing all possible sets of interactions that can be obtained from the four additional time-independent covariates. For all three data sets and both versions of the HCV variable, the preferred model was the same: Besides the interaction between the square root of time and HCV status, it includes both the interactions of the square root of time with baseline AIDS and with duration of NRTI intake before therapy start. Models which included active intravenous drug consumption as additional time-dependent covariate were not preferred by BIC.

Note that the previously induced method by Harrison et al. (2009) requires the inclusion of all possible interactions between the time variable and the other covariates which

would lead to a needlessly large model. This is not only relatively hard to interpret and potentially confusing, but it also does not result in a better model with respect to the BIC. For this reason, we propose an extended correction method that covers both the case of less interactions and the potential inclusion of time-dependent covariates. The latter is not necessary for our univariate models, but would be needed to include HIV-1 RNA if desired.

3 Baseline correction

3.1 Basic correction method

We briefly review the correction method as described in Harrison et al. (2009). Suppose that for an arbitrary individual i with $i = 1, ..., I$, we have $N_i + 1$ measurements $y_{i0}, y_{i1}, y_{i2}..., y_{iN_i}$ at different time points $t_{i0}, t_{i1}, t_{i2}, ..., t_{iN_i}$. Let y_{i0} be a baseline measurement, whereas $y_{i1}, y_{i2}, ..., y_{iN_i}$ are follow-up measurements. Additionally, let potential other time-constant covariates be included in the vector $\boldsymbol{x}_i$. The general model for an arbitrary individual i at some time point t is

$$y_{it} = \beta_0 + t\beta_1 + \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + (\boldsymbol{x}_i t)^T\boldsymbol{\beta}_3 + b_{0i} + tb_{1i} + \epsilon_{it},$$

where y_{it} is the outcome measurement at time t, and $\boldsymbol{x}_i$ denotes the vector of time-invariant covariates. The vector of fixed coefficients is $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$, and $\boldsymbol{b}_i = (b_{0i}, b_{1i})$ is a vector containing the subject-specific random intercept and slope, which are multivariate normally distributed with expected value $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The measurement error is normally distributed $\epsilon_{it} \sim N(0, \xi^2)$.

Now consider the underlying true value $y^*_{it} = y_{it} - \epsilon_{it}$ and the change $d^*_{it} = y^*_{it} - y^*_{i0}$ in the underlying true response value from baseline to time point t.

The expected value of d^*_{it} given the underlying baseline value turns out to be

$$\mathrm{E}(d^*_{it} \,|\, y^*_{i0}) = t(\beta_1 - \beta_0\gamma) + (\boldsymbol{x}_i t)^T(\boldsymbol{\beta}_3 - \boldsymbol{\beta}_2\gamma) + y^*_{i0}t\gamma,$$

with

$$\gamma = \frac{\mathrm{Cov}(b_{0i}, b_{1i})}{\mathrm{Var}(b_{0i})}. \tag{2}$$

This leads to the adjusted parameters

$$\beta_1^A = \beta_1 - \beta_0\gamma \quad \text{and} \quad \boldsymbol{\beta}_3^A = \boldsymbol{\beta}_3 - \boldsymbol{\beta}_2\gamma,$$

which enables the examination of the association between covariates and subsequent progression of the outcome variable for patients with the same baseline value y^*_{i0}.

In practice, estimated coefficients and covariance matrices from the fitted model are used, leading to

$$\hat{\boldsymbol{\beta}}_3^A = \hat{\boldsymbol{\beta}}_3 - \hat{\boldsymbol{\beta}}_2\hat{\gamma} \quad \text{and} \quad \hat{\beta}_1^A = \hat{\beta}_1 - \hat{\beta}_0\hat{\gamma},$$

where

$$\hat{\gamma} = \frac{\widehat{\mathrm{Cov}}(b_{0i}, b_{1i})}{\widehat{\mathrm{Var}}(b_{0i})}.$$

Standard errors of the adjusted parameter estimates can be obtained using the delta method.

3.2 Extension to covariates that do not interact with time, and which may be time-varying

The model described above can only be applied using time-constant covariates, and interactions of all covariates with time have to be included. However, it can be extended so that both of these possibilities are given. A more general model has the form

$$y_{it} = \beta_0 + t\beta_1 + \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + (\boldsymbol{x}_i t)^T\boldsymbol{\beta}_3 + \boldsymbol{w}_i^T\boldsymbol{\beta}_4 + \boldsymbol{c}_{it}^T\boldsymbol{\beta}_5 + b_{0i} + tb_{1i} + \epsilon_{it}.$$

where $\boldsymbol{w}_i$ represents a vector of time-constant covariates that do not interact with time, and $\boldsymbol{c}_{it}$ is a vector with time-dependent covariates. Now consider the underlying true value $y_{it}^* = y_{it} - \epsilon_{it}$ which is for $t = 0$

$$y_{i0}^* = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \boldsymbol{w}_i^T\boldsymbol{\beta}_4 + \boldsymbol{c}_{i0}^T\boldsymbol{\beta}_5 + b_{0i}. \tag{3}$$

The change $d_{it}^* = y_{it}^* - y_{i0}^*$ in the underlying true response value can now be written as

$$d_{it}^* = (\boldsymbol{c}_{it} - \boldsymbol{c}_{i0})^T\boldsymbol{\beta}_5 + t\beta_1 + (\boldsymbol{x}_i t)^T\boldsymbol{\beta}_3 + tb_{1i}.$$

From this we can derive the conditional expectation of d_{it}^* given y_{i0}^*:

$$\mathrm{E}(d_{it}^* \,|\, y_{i0}^*) = (\boldsymbol{c}_{it}-\boldsymbol{c}_{i0})^T\boldsymbol{\beta}_5+t\beta_1+(\boldsymbol{x}_i t)^T\boldsymbol{\beta}_3+\mathrm{E}(tb_{1i} \,|\, b_{0i} = y_{i0}^*-\beta_0-\boldsymbol{x}_i^T\boldsymbol{\beta}_2-\boldsymbol{w}_i^T\boldsymbol{\beta}_4-\boldsymbol{c}_{i0}^T\boldsymbol{\beta}_5).$$

Expression (3) is used to obtain the conditional expected value of tb_{1i}, using the joint multivariate distribution of b_{0i} and b_{1i}:

$$\mathrm{E}(tb_{1i} \,|\, b_{0i} = y_{i0}^*-\beta_0-\boldsymbol{x}_i^T\boldsymbol{\beta}_2-\boldsymbol{w}_i^T\boldsymbol{\beta}_4-\boldsymbol{c}_{i0}^T\boldsymbol{\beta}_5) = (y_{i0}^*-\beta_0-\boldsymbol{x}_i^T\boldsymbol{\beta}_2-\boldsymbol{w}_i^T\boldsymbol{\beta}_4-\boldsymbol{c}_{i0}^T\boldsymbol{\beta}_5)t\gamma, \tag{4}$$

with γ as in (2). This leads to the final form of this conditional expectation:

$$\mathrm{E}(d_{it}^* \,|\, y_{i0}^*) = t(\beta_1-\beta_0\gamma)+(\boldsymbol{x}_i t)^T(\boldsymbol{\beta}_3-\boldsymbol{\beta}_2\gamma)+(y_{i0}^*-\boldsymbol{w}_i^T\boldsymbol{\beta}_4-\boldsymbol{c}_{i0}^T\boldsymbol{\beta}_5)t\gamma+(\boldsymbol{c}_{it}-\boldsymbol{c}_{i0})^T\boldsymbol{\beta}_5. \tag{5}$$

In practice, estimated coefficients and covariance matrices from the fitted model are used, and also the calculation of standard errors, p-values and confidence intervals works as explained above.

As we see, the adjustments β_1^A and $\boldsymbol{\beta}_3^A$ are the same as in the last section. By contrast, the term $y_{i0}^* t\gamma$, including the assumed underlying baseline value, is extended to $(y_{i0}^* - \boldsymbol{w}_i^T\boldsymbol{\beta}_4 - \boldsymbol{c}_{i0}^T\boldsymbol{\beta}_5)t\gamma$. The so far time-constant covariates $\boldsymbol{w}_i$ become time-dependent due to the multiplication with t, and thus influence the course of the expected change over time. The second additional component, $(\boldsymbol{c}_{it} - \boldsymbol{c}_{i0})^T\boldsymbol{\beta}_5$, represents the difference from baseline induced by the time-dependent covariates. Note that a change in a time-dependent covariate - especially if it is categorical - can result in an abrupt alteration in the resulting expected change. The potential influence of these two additional parts in equation (5) is further illustrated in the following application.

3.3 Use with bivariate models

In some applications, joint modelling of two (or even more) variables may be desired, because one influential variable might be on the causal pathway of the other and/or is also measured with error. This can be done within the linear mixed model framework, but the data set has to be restructured (see Doran and Lockwood, 2006, for more details on estimation with `R` or Thiébaut et al., 2002, for `SAS`). When using `R`, the two outcomes of interest have to be stacked in one column, and a set of dummies has to be created by augmenting the existing variables with zeros, so that the respective covariates are estimated for the two outcomes separately. Note that two separate intercepts have to be used now, and two error terms (one for each outcome) have to be taken into account when fitting the model.

For a multivariate model, the adjustment procedure is derived in a similar way as for the univariate one. The joint model for two outcomes A and B with $\boldsymbol{y}_{it} = (y_{it(A)}, y_{it(B)})$ now has the form

$$\boldsymbol{y}_{it} = \boldsymbol{\beta}_0 + t\boldsymbol{\beta}_1 + \boldsymbol{X}_i^T\boldsymbol{\beta}_2 + (\boldsymbol{X}_i t)^T\boldsymbol{\beta}_3 + \boldsymbol{W}_i^T\boldsymbol{\beta}_4 + \boldsymbol{C}_{it}^T\boldsymbol{\beta}_5 + \boldsymbol{b}_{0i} + t\boldsymbol{b}_{1i} + \boldsymbol{\epsilon}_{it}$$

where each element of $\boldsymbol{\beta}$ is now a vector and comprises both the coefficients of outcome A and B stacked above each other, e.g.

$$\boldsymbol{\beta}_0 = \begin{pmatrix} \beta_{0(A)} \\ \beta_{0(B)} \end{pmatrix}.$$

Note that some elements of $\boldsymbol{\beta}$ can also be zero if different covariates for both outcomes are desired. Consequently, the matrices $\boldsymbol{X}_i, \boldsymbol{W}_i$ and $\boldsymbol{C}_{it}$ have two colums with the respective covariate vector and zeros stacked above each other, e.g.

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x}_i \end{pmatrix}.$$

The random effects vectors $\boldsymbol{b}_{0i}^T = (b_{0i(A)}, b_{0i(B)})$ and $\boldsymbol{b}_{1i}^T = (b_{1i(A)}, b_{1i(B)})$ also cover both outcomes, so that $\boldsymbol{b}_i^T = (\boldsymbol{b}_{0i}^T, \boldsymbol{b}_{1i}^T)$ with covariance matrix

$$\begin{aligned}\boldsymbol{\Sigma} &= \begin{pmatrix} \mathrm{Var}(b_{0i(A)}) & \mathrm{Cov}(b_{0i(A)}, b_{0i(B)}) & \mathrm{Cov}(b_{0i(A)}, b_{1i(A)}) & \mathrm{Cov}(b_{0i(A)}, b_{1i(B)}) \\ \mathrm{Cov}(b_{0i(A)}, b_{0i(B)}) & \mathrm{Var}(b_{0i(B)}) & \mathrm{Cov}(b_{0i(B)}, b_{1i(A)}) & \mathrm{Cov}(b_{0i(B)}, b_{1i(B)}) \\ \mathrm{Cov}(b_{0i(A)}, b_{1i(A)}) & \mathrm{Cov}(b_{0i(B)}, b_{1i(A)}) & \mathrm{Var}(b_{1i(A)}) & \mathrm{Cov}(b_{1i(A)}, b_{1i(B)}) \\ \mathrm{Cov}(b_{0i(A)}, b_{1i(B)}) & \mathrm{Cov}(b_{0i(B)}, b_{1i(B)}) & \mathrm{Cov}(b_{1i(A)}, b_{1i(B)}) & \mathrm{Var}(b_{1i(B)}) \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_{01}^T \\ \boldsymbol{\Sigma}_{01} & \boldsymbol{\Sigma}_1 \end{pmatrix},\end{aligned}$$

where each submatrix of $\boldsymbol{\Sigma}$ is of dimension 2×2. Finally, the measurement errors $\boldsymbol{\epsilon}_{it}$ are normally distributed with

$$\boldsymbol{\epsilon} \sim N\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \xi_{(A)}^2 & 0 \\ 0 & \xi_{(B)}^2 \end{bmatrix} \right\}.$$

As before, the change in the underlying true response value is defined as $\boldsymbol{d}^*_{it} = \boldsymbol{y}^*_{it} - \boldsymbol{y}^*_{i0}$. In contrast to the univariate case, the expectation of $\boldsymbol{d}^*_{it}$ is now conditional on the baseline values of both outcome variables $\boldsymbol{y}^{*T}_{i0} = (y^*_{i0(A)}, y^*_{i0(B)})$:

$$\mathrm{E}(\boldsymbol{d}^*_{it} \,|\, \boldsymbol{y}^*_{i0}) = (\boldsymbol{C}_{it} - \boldsymbol{C}_{i0})^T\boldsymbol{\beta}_5 + t\boldsymbol{\beta}_1 + (\boldsymbol{X}_i t)^T\boldsymbol{\beta}_3 + \mathrm{E}(t\boldsymbol{b}_{1i} \,|\, \boldsymbol{y}^*_{i0}).$$

Instead of conditioning on the underlying baseline values $\boldsymbol{y}^*_{i0}$, we condition on the equivalent expression

$$\boldsymbol{b}_{0i} = \boldsymbol{y}^*_{i0} - \boldsymbol{\beta}_0 - \boldsymbol{X}_i^T\boldsymbol{\beta}_2 - \boldsymbol{W}_i^T\boldsymbol{\beta}_4 - \boldsymbol{C}_{i0}^T\boldsymbol{\beta}_5. \tag{6}$$

As in the univariate deduction of formula (4), we make use of the multivariate normality of the random effects terms to derive the elements $\mathrm{E}(t\boldsymbol{b}_{1i} \,|\, \boldsymbol{b}_{0i})$ and obtain

$$\mathrm{E}(t\boldsymbol{b}_{1i} \,|\, \boldsymbol{b}_{0i}) = t\boldsymbol{\Gamma}\boldsymbol{b}_{0i},$$

where

$$\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_0^{-1}.$$

The conditional expected value of $\boldsymbol{d}^*_{it}$ is therefore

$$\mathrm{E}(\boldsymbol{d}^*_{it} \,|\, \boldsymbol{y}^*_{i0}) = (\boldsymbol{C}_{it} - \boldsymbol{C}_{i0})^T\boldsymbol{\beta}_5 + t\boldsymbol{\beta}_1 + (\boldsymbol{X}_i t)^T\boldsymbol{\beta}_3 + t\boldsymbol{\Gamma}\boldsymbol{b}_{0i}.$$

If we have a look at the expected underlying change of only one outcome, e.g. outcome A, the corresponding part of this expression can be written as

$$\begin{aligned}\mathrm{E}(d^*_{it(A)} \,|\, \boldsymbol{y}^*_{i0}) = {} & t(\beta_{1(A)} - \beta_{0(A)}\gamma_{11} - \beta_{0(B)}\gamma_{12}) + (\boldsymbol{x}_i t)^T(\boldsymbol{\beta}_{3(A)} - \boldsymbol{\beta}_{2(A)}\gamma_{11} - \boldsymbol{\beta}_{2(B)}\gamma_{12}) \\ & + (y^*_{i0(A)} - \boldsymbol{w}_i^T\boldsymbol{\beta}_{4(A)} - \boldsymbol{c}_{i0}^T\boldsymbol{\beta}_{5(A)})t\gamma_{11} + (y^*_{i0(B)} - \boldsymbol{w}_i^T\boldsymbol{\beta}_{4(B)} - \boldsymbol{c}_{i0}^T\boldsymbol{\beta}_{5(B)})t\gamma_{12} \\ & + (\boldsymbol{c}_{it} - \boldsymbol{c}_{i0})^T\boldsymbol{\beta}_{5(A)}\end{aligned}$$

where the subscripts (A) and (B) mean that only the part of the coefficient vectors that is related to outcome A or B, respectively, is used, and γ_{11} and γ_{12} are the entries in the upper row of the matrix $\boldsymbol{\Gamma}$. Note that the corrected coefficients have a similar form as in the univariate case (5), but γ_{11} and γ_{12} are more complicated than γ and the underlying baseline value $y^*_{i0(B)}$ of outcome B needs to be incorporated.

4 Application

In this section, the extended method explained above shall be applied to the data sets from the Swiss HIV Cohort Study in order to investigate if patients with differing hepatitis C status show a different slope of the CD4 cell counts over time. We first present univariate models that were selected by BIC as described in Section 2.3 which all do not include HIV-RNA as time-dependent covariate. As mentioned in Section 2.2, the inclusion of this variable is quite controversial, because it is on the causal pathway between HCV status and CD4 count and is subject to additional measurement error. Therefore, we show in Section 4.3 how a bivariate model for two outcomes, namely the square root of CD4 counts and $\log_{10}$ of HIV-1 RNA, can be used to tackle this problem. All calculations were conducted in `R`, using the packages `lme4` and `nlme` for model fitting. The `R` functions used for adjusting the coefficients in the univariate and bivariate models can be found in the supplementary material.

4.1 Interpretation of the selected model

Tables 1 and 2 show the corrected coefficients $\hat{\beta}_1^A$ and $\hat{\boldsymbol{\beta}}_3^A$ that are necessary to compare the expected change from baseline. The first table covers the models with negative versus positive hepatitis C status. In all three data sets, the corrected coefficient of the interaction between the square root of time and HCV status is significant and negative. This means that there is strong evidence that the increase in CD4 cell counts of a person with hepatitis C is slower than in the case of an HCV-negative patient, if they start at the same baseline value. With an estimated value of -1.12, this effect is strongest in data set M, whereas it is clearly smaller in the other two data sets (-0.53 in data set O, -0.72 in data set N).

The same pattern is found concerning the interaction between the square root of time and the duration of previous NRTI intake: the longer a patient remained under therapy before HAART start, the slower the increase in CD4 cell counts. In contrast, the coefficient of the third interaction between the square root of time and AIDS at baseline is positive, suggesting that a person with AIDS would experience a quicker rise of CD4 cell counts. To our knowledge, this has not been reported in other papers so far and is somewhat counter-intuitive, because one would expect a slower rise if the respective person already has AIDS. A potential explanation for this observation may be that patients who already suffer from AIDS are in most cases in a much worse physical condition than persons who have not yet developed AIDS. Thus, they have a much stronger motivation to stick to their HIV treatment, and less adherence problems occur, leading to increased efficacy of the treatment.

Table 2 refers to the models with three categories of hepatitis C status, joint p-values were calculated using a Wald test. In data sets O and N, the corrected coefficients of the interaction between the time variable and HCV status are in the expected order, i.e. the increase for a patient with inactive hepatitis C is closer to the reference category than for a patient in the active category. For data set M, the respective coefficients of both HCV categories are in the reversed order, but very close to each other (-1.16 and -1.11), and considerably lower than in the other two data sets. Only in this case both interaction terms are significant ($p < 0.001$ and $p < 0.001$, respectively). Larger differences between the two HCV categories are found in data sets O (-0.41 and -0.56) and N (-0.48 and -0.80), where at the same time no evidence for differences between negative and inactive hepatitis C status can be detected. In all three data sets, duration of NRTI intake has a significant negative effect, whereas AIDS seems to have a positive effect on the increase of CD4 cell counts.

Table 3 shows the estimated coefficients of the model selected by BIC fitted to data set O before our proposed correction. Note that in contrast to the corrected coefficients in Table 2, the p-value of the uncorrected interaction coefficient between the square root of time and active hepatitis C ($p = 0.680$) and the joint p-value from a Wald test of both interaction terms ($p = 0.915$) are not significant. This shows that omitting the necessary baseline correction would bias the answer to our main research question, a situation that occurs in five of the six models presented here.

The differences between the HCV categories in all three data sets are reflected in

Table 1: Coefficients of the corrected selected model for the change from baseline CD4 cell count with binary hepatitis C status

	Value	Std.Error	CI, lower	CI, upper	p-value
Data set O:					
sqrt(time)	8.27	0.15	7.98	8.57	< 0.001
sqrt(time) x HCV (pos.)	-0.53	0.13	-0.79	-0.28	< 0.001
sqrt(time) x AIDS (yes)	0.35	0.14	0.07	0.64	0.013
sqrt(time) x NRTI duration	-0.19	0.03	-0.25	-0.12	< 0.001
Data set M:					
sqrt(time)	9.06	0.19	8.70	9.43	< 0.001
sqrt(time) x HCV (pos.)	-1.12	0.17	-1.45	-0.80	< 0.001
sqrt(time) x AIDS (yes)	0.36	0.18	0.01	0.71	0.042
sqrt(time) x NRTI duration	-0.10	0.03	-0.16	-0.03	0.003
Data set N:					
sqrt(time)	8.56	0.12	8.33	8.79	< 0.001
sqrt(time) x HCV (pos.)	-0.72	0.17	-1.06	-0.38	< 0.001
sqrt(time) x AIDS (yes)	0.63	0.19	0.25	1.00	0.001
sqrt(time) x NRTI duration	-0.09	0.03	-0.14	-0.04	0.001

Table 2: Coefficients of the corrected selected model for the change from baseline CD4 cell count with three categories of hepatitis C status

	Value	Std.Error	CI, lower	CI, upper	p-value	joint p-value
Data set O:						
sqrt(time)	8.28	0.15	7.98	8.57	< 0.001	-
sqrt(time) x HCV (inact.)	-0.41	0.26	-0.93	0.11	0.120	< 0.001
sqrt(time) x HCV (act.)	-0.56	0.14	-0.83	-0.29	< 0.001	
sqrt(time) x AIDS (yes)	0.35	0.14	0.07	0.63	0.014	-
sqrt(time) x NRTI duration	-0.19	0.03	-0.25	-0.12	< 0.001	-
Data set M:						
sqrt(time)	9.06	0.19	8.70	9.43	< 0.001	-
sqrt(time) x HCV (inact.)	-1.16	0.32	-1.79	-0.53	< 0.001	< 0.001
sqrt(time) x HCV (act.)	-1.11	0.18	-1.47	-0.75	< 0.001	
sqrt(time) x AIDS (yes)	0.36	0.18	0.01	0.71	0.042	-
sqrt(time) x NRTI duration	-0.10	0.03	-0.16	-0.03	0.003	-
Data set N:						
sqrt(time)	8.56	0.12	8.33	8.79	< 0.001	-
sqrt(time) x HCV (inact.)	-0.48	0.33	-1.13	0.16	0.140	< 0.001
sqrt(time) x HCV (act.)	-0.80	0.20	-1.19	-0.41	< 0.001	
sqrt(time) x AIDS (yes)	0.63	0.19	0.25	1.00	< 0.001	-
sqrt(time) x NRTI duration	-0.09	0.03	-0.14	-0.03	0.001	-

	Value	Std. Error	p-value	joint p-value
Intercept	19.84	0.52	<0.001	-
sqrt(time)	3.19	0.10	<0.001	-
HCV (inact.)	-1.42	0.60	0.018	< 0.001
HCV (act.)	-1.92	0.32	<0.001	
AIDS (yes)	-6.54	0.32	<0.001	-
NRTI duration	-0.08	0.08	0.298	-
sex (female)	0.38	0.25	0.136	-
age (per year)	-0.06	0.01	<0.001	-
sqrt(time) x HCV (inact.)	-0.04	0.32	0.888	0.915
sqrt(time) x HCV (act.)	-0.07	0.17	0.680	
sqrt(time) x AIDS (yes)	2.03	0.17	<0.001	-
sqrt(time) x NRTI duration	-0.17	0.04	<0.001	-

Table 3: Selected model for data set O with three HCV categories

Figure 2 which exemplifies the expected change of the CD4 cell counts of the different groups for a common baseline square root of the CD4 cell counts of 16. The values of the other covariates are chosen to be the reference category in the case of binary variables (no AIDS at baseline, male gender) and the mean value in the respective data set in the case of continuous variables. We can see that the course of the CD4 cell counts of patients with active and inactive hepatitis C is so similar in the middle data set that the two respective lines cannot be distinguished, whereas differences can be seen in the two other data sets. One would expect, that the CD4 cell counts recover faster in persons with inactive than with active HCV status, a behaviour which is reflected best in the results from the newest data set. If this figure is compared to the moving average in Figure 1, we see that the course of the CD4 cell counts is comparable in the case of data set M, however, the inclusion of covariates and the appropriate baseline correction lead to strong differences in the case of the other two data sets.

4.2 Impact of different covariate values

We also assess the consequences of different choices of the covariate values, especially the newly introduced $\boldsymbol{w}_i$. In the upper left of Figure 3, we see the expected change of the square root of the CD4 cell counts for a baseline value of 16, if all covariates are chosen as mean values and reference categories, respectively, as has been done so far in this paper. Replacing the covariate values by different values can sometimes have considerable impact on the predicted course on the square root of the CD4 counts, which apart from the estimated coefficient also depends on the nature of the respective covariate, i.e. if it is interacting with the time variable.

In the present figure, we show three different choices of particular covariate values and compare the predicted courses of the square root of CD4 counts resulting from that choice to the one from the originally used covariate combination. A change of the

Table 4: Corrected coefficients of the CD4 part of the bivariate models for CD4 counts and HIV-1 RNA with two HCV categories

	Value	Std.Error	CI, lower	CI, upper	p-value
Data set O:					
sqrt(time)	7.59	0.13	7.34	7.85	< 0.001
sqrt(time) x HCV (pos.)	-0.47	0.13	-0.73	-0.21	< 0.001
sqrt(time) x AIDS (yes)	0.74	0.15	0.46	1.03	< 0.001
sqrt(time) x NRTI duration	-0.18	0.03	-0.25	-0.12	< 0.001
Data set M:					
sqrt(time)	6.26	0.16	5.95	6.56	< 0.001
sqrt(time) x HCV (pos.)	-1.00	0.17	-1.34	-0.67	< 0.001
sqrt(time) x AIDS (yes)	0.72	0.18	0.37	1.08	< 0.001
sqrt(time) x NRTI duration	-0.11	0.03	-0.18	-0.05	< 0.001
Data set N:					
sqrt(time)	3.24	0.11	3.02	3.46	< 0.001
sqrt(time) x HCV (pos.)	-0.41	0.18	-0.76	-0.07	0.019
sqrt(time) x AIDS (yes)	0.94	0.19	0.57	1.32	< 0.001
sqrt(time) x NRTI duration	-0.08	0.03	-0.13	-0.02	0.0059

variable AIDS at baseline (which also interacts with time) from 0 to 1 is seen in the upper right and results in a quicker increase in CD4 cell counts over time, analogous to the positive corrected coefficient in Table 2. If we instead change the value of age, a covariate that does not interact with the square root of time, from the mean value 40.3 to 60 years and leave all other covariates unchanged, the resulting curves are markedly less steep than before, as we can see in the lower left plot.

Finally, the plot in the lower right corner of Figure 3 shows the predicted change in square root of CD4 cell counts when the duration of NRTI intake before start of highly active antiretroviral therapy is changed from 0.5 years to 5 years. We can see that the increase in CD4 cell counts is only very slightly slower, again in agreement with the corrected coefficients in Table 2.

4.3 Bivariate modelling of CD4 and HIV-1 RNA

As mentioned before, it would be desirable to account for $\log_{10}$ HIV-1 RNA in the univariate models. This could be easily done by including this variable in the models as time-dependent covariate, however, this approach is associated with several problems as described towards the end of Section 2.2. For demonstration, we fitted one model including $\log_{10}$ HIV-1 RNA and three categories of hepatitis C to data set O, but the interpretation of the coefficients is difficult and has to be done with caution:

Table 5: Corrected coefficients of the CD4 part of the bivariate models for CD4 counts and HIV-1 RNA with three HCV categories

	Value	Std.Error	CI, lower	CI, upper	p-value	joint p-value
Data set O:						
sqrt(time)	7.61	0.13	7.35	7.86	< 0.001	-
sqrt(time) x HCV (inact.)	-0.35	0.27	-0.87	0.18	0.200	0.002
sqrt(time) x HCV (act.)	-0.49	0.14	-0.77	-0.22	< 0.001	
sqrt(time) x AIDS (yes)	0.75	0.15	0.47	1.04	< 0.001	-
sqrt(time) x NRTI duration	-0.18	0.03	-0.25	-0.12	< 0.001	-
Data set M:						
sqrt(time)	6.26	0.16	5.96	6.57	< 0.001	-
sqrt(time) x HCV (inact.)	-1.06	0.33	-1.70	-0.42	0.0012	< 0.001
sqrt(time) x HCV (act.)	-0.99	0.19	-1.36	-0.62	< 0.001	
sqrt(time) x AIDS (yes)	0.72	0.18	0.37	1.08	< 0.001	-
sqrt(time) x NRTI duration	-0.11	0.03	-0.18	-0.05	< 0.001	-
Data set N:						
sqrt(time)	3.23	0.11	3.01	3.45	< 0.001	-
sqrt(time) x HCV (inact.)	-0.12	0.33	-0.77	0.53	0.720	0.036
sqrt(time) x HCV (act.)	-0.51	0.20	-0.91	-0.12	0.01	
sqrt(time) x AIDS (yes)	0.95	0.19	0.57	1.32	< 0.001	-
sqrt(time) x NRTI duration	-0.08	0.03	-0.13	-0.02	0.006	-

The estimated coefficient of $\log_{10}$ HIV-1 RNA is -0.71, while the other coefficients remain relatively similar to the model without HIV-1 RNA. If the expected change from baseline is to be calculated for several time points as in Section 4.2, assumptions on the associated HIV-1 RNA values have to be made, which turns out to be difficult for such a time-dependent variable: Assuming a constant value for all time points would be unrealistic, whereas assuming a frequently changing course of HIV-1 RNA might lead to unnecessarily confusing plots.

To solve the problems that occur when HIV-1 RNA is included as covariate, a bivariate model for CD4 counts and HIV-1 RNA with the same covariates as in the univariate model selected by BIC is fitted. After that, the bivariate correction method is applied. Note that we now condition on both the underlying baseline CD4 count and the underlying baseline HIV-RNA. In Tables 4 and 5 we see the results of the correction for the CD4 part of the model (the results for HIV-1 RNA are omitted because they are not of primary interest). As before, an infection with HCV causes a significantly slower increase in the square root of CD4 counts in all data sets, and this effect is comparable to the results in the univariate case for data sets O and M (data set O: -0.47, data set M: -1.00), and slightly less pronounced than in the univariate case in data set N (-0.41). The coefficients for three categories of HCV infection are also comparable to the former results, but again we see an effect that is slightly weaker than before in data set N.

A limiting factor in our bivariate analyses is the nature of our HIV-1 RNA measurements: Many persons show relatively high HIV-1 RNA measurements at baseline, but the following HIV treatment normally causes the HIV-1 RNA to fall below a certain detection limit, so that it cannot be determined exactly any more. To avoid a huge number of missing values, some imputation strategy has to be applied for replacing these left-censored values. An overview of possible techniques for that task, especially with longitudinal data, can be found in Jacqmin-Gadda et al. (2000). A newer semiparametric approach is presented in Vock et al. (2012). To keep our analysis as simple as possible, we use half the respective detection limit when an actual observation is missing, although this might introduce some bias. From this follows that $\log_{10}$ of HIV-1 RNA is not necessarily normally distributed any more, which can be problematic if it is intended as outcome variable instead of being a covariate. Despite that we think that the application of a bivariate model and our proposed methodology are justifiable in our case: Even with the imputed HIV-1 RNA values, the histogram of $\log_{10}$ of HIV-1 RNA follows more or less a normal distribution. Besides, there are only very few missing values at baseline, and only they play a direct role in our approach. As those values are almost complete, the proposed adjustment method should remain valid.

4.4 Comparison with original model

In this subsection, the results obtained using the original method by Harrison et al. (2009) are compared and contrasted with the results from our proposed extensions. Figure 4 shows four plots from the new data set in the case of three categories of hepatitis C and an assumed underlying baseline value of 16: The two plots in the upper row show results from models using the original procedure as suggested by Harrison et al. (2009). The

left plot depicts the expected change in CD4 cell counts from a univariate model with all possible interactions with the time variable, but without HIV-1 RNA. This variable and its interaction with the square root of time have to be left out due to the fact that time-dependent covariates cannot be taken into account in the original procedure. As an alternative to this model without HIV-1 RNA, the upper right plot depicts a model where the decimal logarithm of the baseline HIV-1 RNA measurement is included as time-constant covariate, as well as its interaction with the time variable. The values of the covariates are chosen as in Figure 2, the mean of baseline $\log_{10}$ HIV-1 RNA is 4.35. Note again that the interpretation of $\log_{10}$ HIV-1 RNA as time-dependent covariate is difficult, for which reason we do not recommend to include it, but to use a bivariate model with CD4 cell counts and $\log_{10}$ HIV-1 RNA as outcome variables instead. Both models in the upper row have a considerably larger BIC than the model selected by BIC.

In the lower row on the left, we see the expected change of the CD4 cell counts from the model selected by BIC (i.e. including HIV-1 RNA as time-dependent covariate) with the proposed correction. On the right, the expected change obtained from the bivariate model with correction is depicted for a $\log_{10}$ HIV-1 RNA baseline value of 4.35, the same as in the upper plot. Between the two models not using HIV-1 RNA (in the left column) we only find small differences, but we can see that the slopes in the univariate model with baseline $\log_{10}$ HIV-1 RNA are slightly less steep than in the other univariate models. Obviously, having to remove an important time-dependent covariate can result in differences in the expected change. The plot obtained from the bivariate model, however, shows a markedly steeper slope than the other three plots, which confirms again that the correct way of treating a covariate with measurement error is important. It has to be mentioned again that in the bivariate case, the depicted trajectories are to be interpreted conditional on the underlying baseline values of both CD4 and $\log_{10}$ HIV-1 RNA. Thus, the trajectory can vary considerably in dependence on the choice of both baseline CD4 cell counts and baseline $\log_{10}$ HIV-1 RNA.

Concerning the univariate models, using more interaction terms than necessary leads to larger standard errors, an effect that is in our case especially pronounced for the corrected coefficient of the time variable $\hat{\beta}_2^A$, where the standard errors of the two models with all possible interactions are about twice as large as in the model selected by BIC. Therefore we conclude that both model extensions would not add much value but result in less accuracy and potentially misleading predictions, and our proposed methodology permits the use of simpler models without losing the possibility to condition on the underlying true baseline value.

5 Summary and outlook

In this paper, we have presented an extension of a method by Harrison et al. (2009) that can be used to take into account the underlying baseline values when comparing the development over time of two or more groups in linear mixed-effects models. We have stressed the importance of conditioning not only on the observed but on the non-observable underlying baseline values. Our extension makes it possible to use both

time-constant and time-dependent covariates. Besides, not all possible interactions with time have to be included, but any subset of these interactions can be used. The influence of the choice of time-dependent, but also of the time-constant covariates on the change over time has been illustrated. Additionally, we have adapted the method for use with bivariate models.

If the main interest lies not in the comparison of several groups, but in obtaining predictions for a concrete individual with certain baseline characteristics and an observed baseline value, this can be also done, however, again with additional terms for time-dependent covariates and covariates without interaction term. The expected change from baseline is obtained by conditioning on the observed rather than the underlying baseline value, as shown in Section 3.6. of Harrison et al. (2009).

In our application, models that took into account only three interaction terms were preferred to models with all possible interactions, using a version of the marginal BIC that can be applied with linear mixed-effects models for model choice. We saw that an HCV-coinfection lead to a significantly slower increase in CD4 cell counts in all three data sets when compared to HCV-negative patients. This was also the case for patients with active hepatitis C status, however, a difference between negative and inactive patients could only be found in one of the three data sets. Similar results were found using a bivariate model for CD4 cell counts and HIV-1 RNA.

Other possible models could incorporate as well time as its square root, probably leading to an even better model fit. However, this combination makes it harder to give a concrete answer to the question if the course of the CD4 cell counts over time is different for groups with differing HCV status. This is the case because the corrected coefficients of both the interactions with time and its square root are needed to answer that question, so that their combined impact on the slope depends on the time point and changes over time.

Acknowledgements:

We thank an associate editor and two referees for their valuable comments on earlier versions of that manuscript. This study has been financed in the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation (grant # 33CS30_134277).

References

Braun, J., Held, L. and Ledergerber, B. (2012). Predictive cross-validation for the choice of linear mixed-effects models with application to data from the Swiss HIV Cohort Study, *Biometrics* **68**(1): 53–61.

Chambless, L. and Davis, V. (2003). Analysis of associations with change in a multivariate outcome variable when baseline is subject to measurement error, *Statistics in Medicine* **22**(7): 1041–1067.

Chambless, L. and Roeback, J. (1993). Methods for assessing difference between groups in change when initial measurement is subject to intra-individual variation, *Statistics in Medicine* **12**(13): 1213–1237.

Cheng, D., Nunes, D., Libmann, H., Vidaver, J., Alperen, J., Saitz, R. and Samet, J. (2007). Impact of hepatitis C on HIV progression in adults with alcohol problems, *Alcoholism: Clinical and Experimental Research* **31**(5): 829–836.

Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd edn, Oxford University Press.

Doran, H. and Lockwood, J. (2006). Fitting value-added models in R, *Journal of Educational and Behavioral Statistics* **31**: 205–230.

Egger, S., Petoumenos, K., Kamarulzaman, A., Hoy, J., Sungkanuparph, S., Chuah, J., Falster, K., Zhou, J. and Law, M. (2009). Long-term patterns in CD4 response is determined by an interaction between baseline CD4 cell count, viral load and time: the Asia Pacific HIV Observational database (APHOD), *Journal of Acquired Immune Deficiency Syndromes* **50**(5): 513–520.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*, Wiley.

Frees, E. W. (2004). *Longitudinal and Panel Data*, Cambridge University Press.

Greub, G., Ledergerber, B., Battegay, M., Grob, P., Perrin, L., Furrer, H., Burgisser, P., Erb, P., Boggian, K., Piffaretti, J., Hirschel, B., Janin, P., Francioli, P., Flepp, M., Telenti, A. and for the Swiss HIV Cohort Study (2000). Clinical progression, survival, and immune recovery during antiretroviral therapy in patients with HIV-1 and hepatitis C virus coinfection: the Swiss HIV Cohort Study, *The Lancet* **356**: 1800–1805.

Harrison, L., Dunn, D., Green, H. and Copas, A. (2009). Modelling the association between patient characteristics and the change over time in a disease measure using observational cohort data, *Statistics in Medicine* **28**: 3260–3275.

Jacqmin-Gadda, H., Thiébaut, R., Chêne, G. and Commenges, D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection, *Biostatistics* **1**(4): 355–368.

Mbougua, J., Laurent, C., Kouanfack, C., Bourgeois, A., Ciaffi, L., Calmy, A., Gwet, H., Koulla-Shiro, S., Ducos, J., Mpoudi-Ngole, E., Molinari, N. and Delaporte, E. (2010). Hepatotoxicity and effectiveness of a Nevirapine-based antiretroviral therapy in HIV-infected patients with or without viral hepatitis B or C infection in Cameroon, *BMC Public Health* **10**: 105–115.

Pauler, D. (1998). The Schwarz criterion and related methods for normal linear models, *Biometrika* **85**(1): 13–27.

Pauler, D., Wakefield, J. and Kass, R. (1999). Bayes factors and approximations for variance component models, *Journal of the American Statistical Association* **94**(448): 1242–1253.

Peters, L., Mocroft, A., Soriano, V., Rockstroh, J., Losso, M., Valerio, L., Aldins, P., Reiss, P., Ledergerber, B., Lundgren, J. D. and for the EuroSIDA Study Group (2009). Hepatitis C virus coinfection does not influence the CD4 cell recovery in HIV-1-infected patients with maximum virologic suppression, *Journal of Acquired Immune Deficiency Syndromes* **50**(5): 457–463.

Pinheiro, J. and Bates, D. (2004). *Mixed-Effects Models in S and S-PLUS*, Springer.

Robbins, G., Daniels, B., Zheng, H., Chueh, H., Meigs, J. and Freedberg, K. (2007). Predictors of antiretroviral treatment failure in an urban HIV clinic, *Journal of Acquired Immune Deficiency Syndromes* **44**(1): 30–37.

Rockstroh, J., Mocroft, A., Soriano, V., Tural, C., Losso, M., Horban, A., Kirk, O., Philips, A., Ledergerber, B., Lundgren, J. and for the EuroSIDA Study Group (2005). Influence of hepatitis C virus infection on HIV-1 disease progression and response to highly active antiretroviral therapy, *The Journal of Infectious Diseases* **192**: 992–1002.

Samet, J., Cheng, D., Libmann, H., Nunes, D., Alperen, J. and Saitz, R. (2007). Alcohol consumption and HIV disease progression, *Journal of Acquired Immune Deficiency Syndromes* **46**(2): 194–199.

Thiébaut, R., Jacqmin-Gadda, H., Chêne, G., Leport, C. and Commenges, D. (2002). Bivariate linear mixed models using SAS proc MIXED, *Computer Methods and Programs in Biomedicine* **69**(3): 249–256.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models, *Biometrika* **92**(2): 351–370.

Verbeke, G., Fieuws, S., Lesaffre, E., Kato, B., Foreman, M., Broos, P. and Milisen, K. (2006). A comparison of procedures to correct for baseline differences in the analysis of continuous longitudinal data: A case study, *Applied Statistics* **55**: 92–101.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer.

Vickers, A. and Altman, D. (2001). Analysing controlled trials with baseline and follow up measurements, *British Medical Journal* **323**: 1123–1124.

Vock, D. M., Davidian, M. and Tsiatis, A. A. (2012). Mixed model analysis of censored longitudinal data with flexible random-effects density, *Biostatistics* **13**(1): 61–73.

Walter, S., Forbes, A., Chan, S., Macaskill, P. and Irwig, L. (2011). When should one adjust for measurement error in baseline variables in observational studies?, *Biometrical Journal* **53**(1): 28–39.

Yanez III, N., Kronmal, R. and Shemanski, L. (1998). The effects of measurement error in response variables and tests of association of explanatory variables in change models, *Statistics in Medicine* **17**(22): 2597–2606.

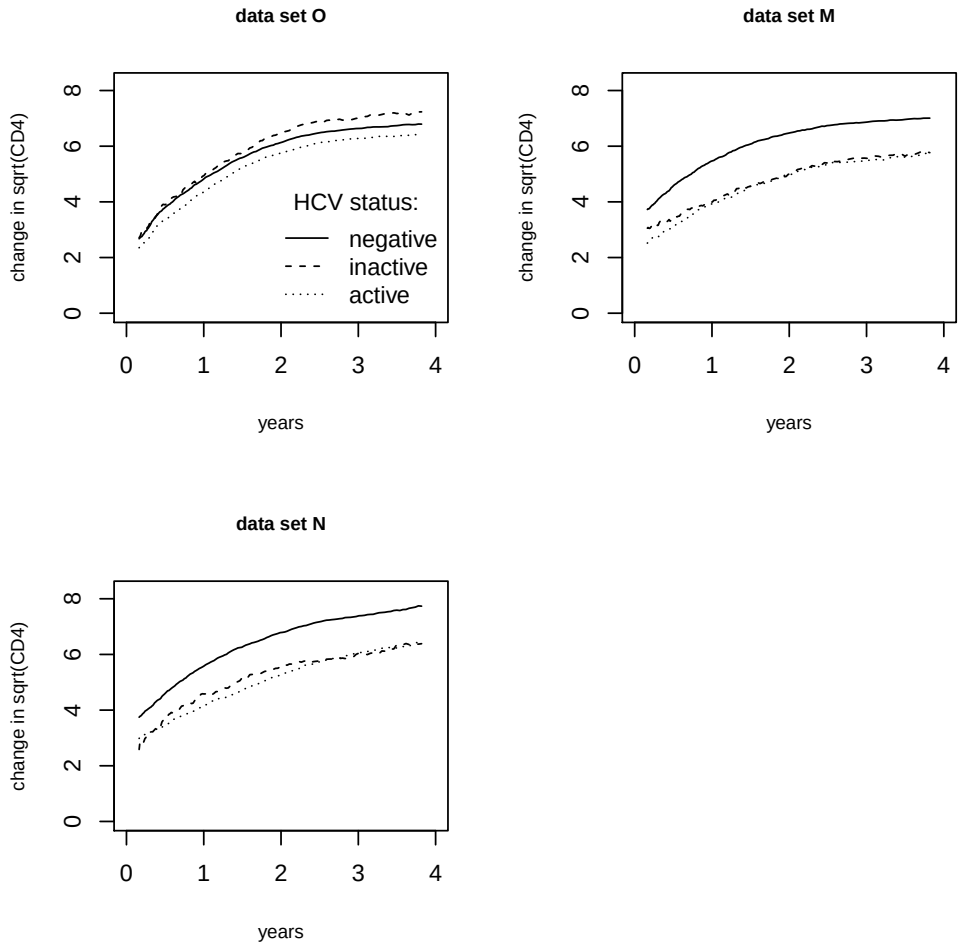

Figure 1: Moving average of the change in square root of the CD4 cell counts for different hepatitis C states in the three data sets (bandwidth 4 months, first time point plotted at 2 months)

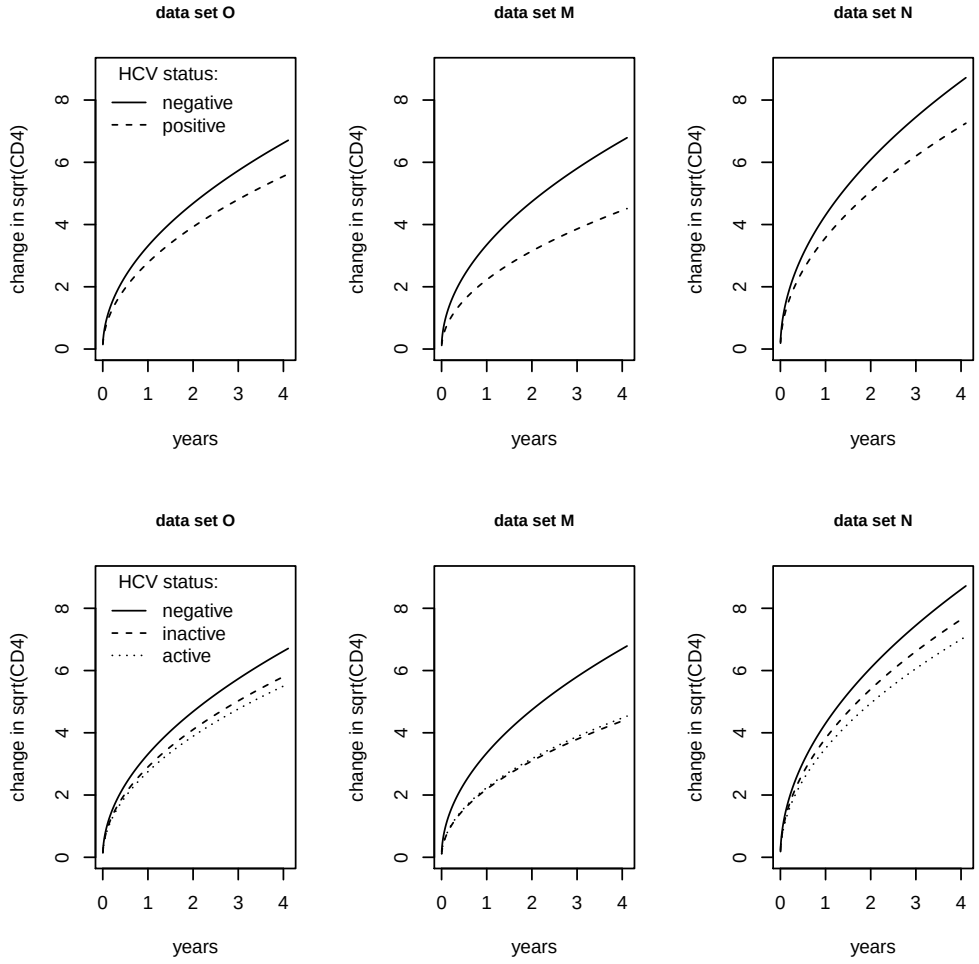

Figure 2: Predicted sqrt(CD4) values for an assumed underlying baseline value of 16

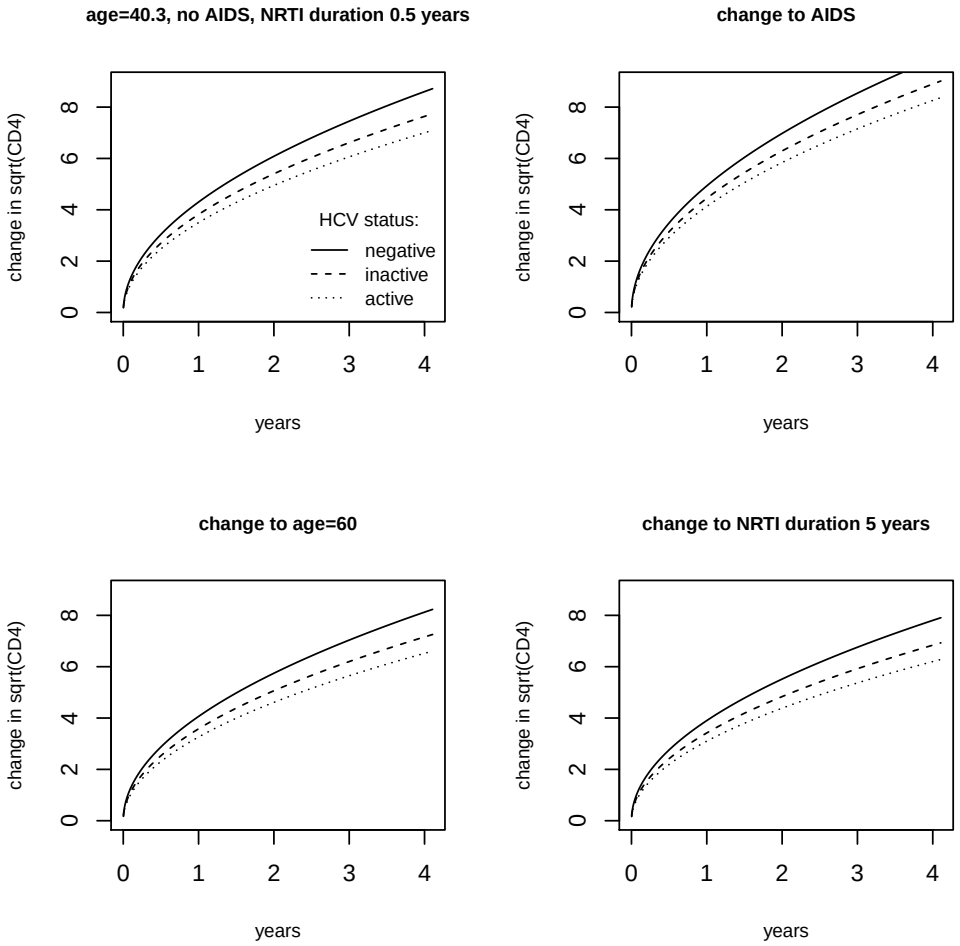

Figure 3: Predicted sqrt(CD4) values for different choices of covariates and an assumed underlying baseline value of 16 and data set N

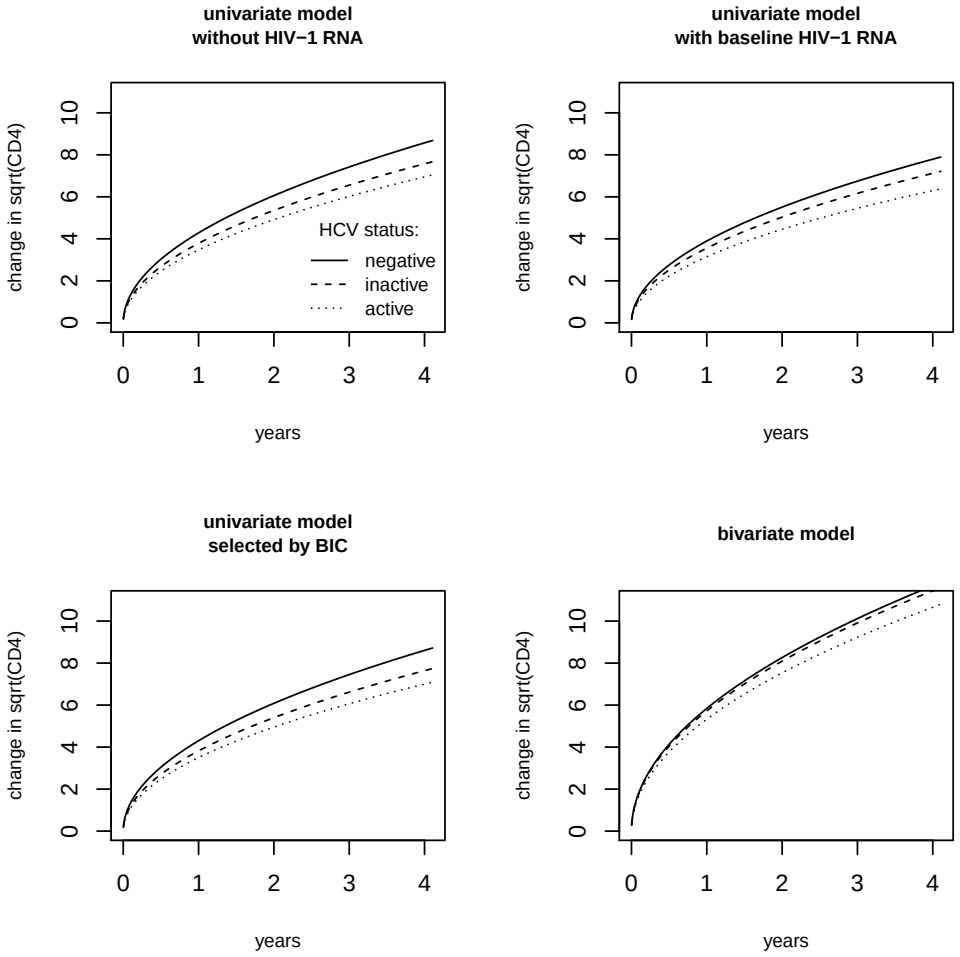

Figure 4: Predicted sqrt(CD4) values for an assumed underlying CD4 cell baseline value of 16, baseline $\log_{10}$ of HIV-1 RNA value of 3 and data set N

Supplementary material: Accounting for baseline differences and measurement error in the analysis of change over time

Julia Braun, Leonhard Held, Bruno Ledergerber
and the Swiss HIV Cohort Study

October 25, 2012

1 General description

This is a short description of the R functions that can be used for the coefficient adjustment as described in the above named article. These programs were developed using R, version 2.14.1, on a laptop with Mac OS X, version 10.5.8, using a 2.4 GHz Intel Core 2 Duo processor.

2 Function `harrison`

This function can be used for linear mixed models with random intercept and slope as described in the paper. These models can be fitted using the `lmer` command from the R package `lme4`. The function just needs the desired model and returns the adjusted coefficients along with their standard errors, confidence intervals and p-values.

Usage:

```
> model <- lmer(outcome ~ (time | id) + time + var1 + var2 + var3 + var1:time +
+     var2:time)
> harrison(model, level = 0.95)
```

Arguments:

- `model`: A fitted `lmer` model with random intercept and slope and interaction terms with time.
- `level`: The desired level of significance of the confidence intervals; default 0.95.

Values:

- `results`: A matrix containing the corrected coefficients, their standard errors, confidence intervals and p-values.
- `cov`: The covariance matrix of the corrected interaction coefficients and the remaining coefficients of the model.

Functions used: `deltamethod` from package `msm`.

3 Function `harrison.bi`

This function can be used for bivariate linear mixed models with random intercept and slope as described in the paper. It needs the desired bivariate model as well as the information, which of the two outcomes is of primary interest and returns the adjusted coefficients of the outcome of interest along with their standard errors, confidence intervals and p-values.

Note that the bivariate model has to have the correct form for use with this function, as explained in the paper. The name of the covariates should make clear for which outcome they are intended, e.g. by adding `.outcome1` after each variable name. The order of the random effects is also important, they should be in the following order: 1) random intercept of variable of interest, 2) random slope of variable of interest, 3) random intercept of second outcome, 4) random slope of second outcome.

As separate error terms for each outcome have to be calculated, this kind of model can only be fitted using the `lme` command from the `R` package `nlme`.

Usage:

```
> model <- lme(fixed = response ~ -1 + dum.out1 + time.out1 + var1.out1 +
+     var2.out1 + var3.out1 + dum.out2 + time.out2 + var1.out2 + var1.out2 +
+     var3.out2 + var1.out1:time.out1 + var2.out1:time.out1 + var1.out2:time.out2 +
+     var2.out2:time.out2, random = ~-1 + dum.out1 + time.out1 + dum.out2 +
+     time.out2 | id, weights = varIdent(form = ~1 | dum.out1))
> harrison.bi(model, level = 0.95, interest = "out1")
```

Arguments:

- `model`: A fitted bivariate `lmer` model with random intercept and slope and interaction terms with time.
- `level`: The desired level of significance of the confidence intervals; default 0.95.
- `interest`: The name of the outcome of interest as used in the names of the respective covariates.

Values:

- `results`: A matrix containing the corrected coefficients, their standard errors, confidence intervals and p-values.
- `cov`: The covariance matrix of the corrected interaction coefficients and the remaining coefficients of the model.

Functions used: `deltamethod` from package `msm`.